

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

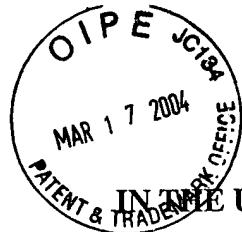
Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

---

Appellant(s): D. Wade Walke, *et al.* Group Art Unit: 1652

Application No.: 09/833,782 Examiner: M.A. Walicka

Filed: 4/12/2001

Title: Novel Human Metalloprotease and Atty. Docket No. LEX-0161-USA  
Polynucleotides Encoding the Same

---

**APPEAL BRIEF**

Mail Stop Appeal Brief  
Commissioner for Patents  
Alexandria, VA 22313

## TABLE OF CONTENTS

I.	REAL PARTY IN INTEREST .....	1
II.	RELATED APPEALS AND INTERFERENCES .....	1
III.	STATUS OF THE CLAIMS .....	2-3
IV.	STATUS OF THE AMENDMENTS .....	3
V.	SUMMARY OF THE INVENTION .....	3
VI.	ISSUES ON APPEAL .....	4
VII.	GROUPING OF THE CLAIMS .....	4
VIII.	ARGUMENT .....	4-18
	A.    Do Claims 1-5 Lack a Patentable Utility? .....	4-18
	B.    Are Claims 1-5 Unusable Due to a Lack of Patentable Utility? .....	18
IX.	APPENDIX .....	19
X.	CONCLUSION .....	20

## TABLE OF AUTHORITIES

### CASES

<i>Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.</i> , 927 F.2d 1200, 18 USPQ2d 1016 (Fed. Cir. 1991) .....	16
<i>Brooktree Corp. v. Advanced Micro Devices, Inc.</i> , 977 F.2d 1555, 1571, 24 USPQ2d 1401 (Fed. Cir. 1992) .....	15
<i>Cross v. Iizuka</i> , 753 F.2d 1040, 224 USPQ 739 (Fed. Cir. 1985) .....	15
<i>Diamond vs. Chakrabarty</i> , 447 U.S. 303, 206 USPQ 193 (U.S., 1980) .....	15
<i>Envirotech Corp. v. Al George, Inc.</i> , 221 USPQ 473, 480 (Fed. Cir. 1984) .....	13
<i>Hoffman v. Klaus</i> , 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988) .....	13
<i>In re Angstadt and Griffin</i> , 537 F.2d 498, 190 USPQ 214 (CCPA 1976) .....	16
<i>In re Brana</i> , 51 F.3d 1560, 34 USPQ2d 1436 (Fed. Cir. 1995) .....	15, 16, 18
<i>In re Fouche</i> , 439 F.2d 1237, 1243, 169 USPQ 429, 434 (CCPA 1971) .....	18
<i>In re Gottlieb</i> , 328 F.2d 1016, 140 USPQ 665 (CCPA 1964) .....	13
<i>In re Jolles</i> , 628 F.2d 1322, 1326 n.11, 206 USPQ 885, 889 n.11 (CCPA 1980) .....	18
<i>In re Langer</i> , 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974) .....	11

<i>In re Marzocchi</i> , 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971) .....	11
<i>In re Malachowski</i> , 530 F.2d 1402, 189 USPQ 432 (CCPA 1976) .....	13
<i>In re Wands</i> , 858 F.2d 731, 8 USPQ 2d 1400 (Fed. Cir. 1988) .....	12, 16
<i>Juicy Whip Inc. v. Orange Bang Inc.</i> , 185 F.3d 1364, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing <i>Brenner v. Manson</i> , 383 U.S. 519, 534 (1966)) .....	15
<i>Raytheon Co. v. Roper Corp.</i> , 724 F.2d 951, 220 USPQ 592 (Fed. Cir. 1983) .....	13
<i>State Street Bank &amp; Trust Co. v. Signature Financial Group Inc.</i> , 149 F.3d 1368, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998) .....	15
<i>Carl Zeiss Stiftung v. Renishaw PLC</i> , 20 USPQ2d 1101 (Fed. Cir. 1991) .....	12

**STATUTES**

35 U.S.C. § 101 ..... 1-2, 4-6, 9, 11-13, 15-18

35 U.S.C. § 112 ..... 2, 4, 5-6, 16-18



## APPEAL BRIEF

Sir:

Appellants hereby submit an original and two copies of this Appeal Brief to the Board of Patent Appeals and Interferences ("the Board") in response to the Final Office Action mailed January 24, 2003. The Notice of Appeal was timely submitted on April 24, 2003, and was received in the Patent and Trademark Office ("the Office") on April 29, 2003. This Appeal Brief is timely submitted in light of the concurrently filed Petition for an Extension of Time of three months to and including September 29, 2003 and authorization to deduct the fee as required under 37 C.F.R. § 1.17(a)(2) from Appellants' Representatives' deposit account. The Commissioner is also authorized to charge the fee for filing this Appeal Brief (\$160.00), as required under 37 C.F.R. § 1.17(c), to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

Appellants believe no fees in addition to the fee for filing the Appeal Brief and the fee for the extension of time are due in connection with this Appeal Brief. However, should any additional fees under 37 C.F.R. §§ 1.16 to 1.21 be required for any reason related to this communication, the Commissioner is authorized to charge any underpayment or credit any overpayment to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

### **I. REAL PARTY IN INTEREST**

The real party in interest is the Assignee, Lexicon Genetics Incorporated, 8800 Technology Forest Place, The Woodlands, Texas, 77381.

### **II. RELATED APPEALS AND INTERFERENCES**

Appellants know of no related appeals or interferences.

### **III. STATUS OF THE CLAIMS**

The present application was filed on April 12, 2001, claiming the benefit of U.S. Provisional Application Number 60/196,319, which was filed on April 12, 2000, and included original claims 1-3. A First Official Action, was issued on January 18, 2002 ("the First Action"),

in which claims 1-3 were rejected under 35 U.S.C. § 101, as allegedly lacking patentable utility and under 35 U.S.C. § 112, first paragraph, as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility, Claim 2 was also rejected under 35 U.S.C. § 112, second paragraph, as allegedly indefinite, claim 3 was rejected under 35 U.S.C. § 102(b), as allegedly being anticipated. In a response to the First Official Action, submitted to the Office on April 18, 2002 ("response to the First Action"), Appellants amended claims 1-2 to further improve their clarity, and respectfully traversed the rejections of claim 1-3 under 35 U.S.C. § 101 and claim 1 under 35 U.S.C. § 112, first paragraph.

A Second Official Action, was issued on July 3, 2002 (the "Second Action"), in which rejection of claims 1-3 was maintained under 35 U.S.C. § 101 as allegedly lacking a patentable utility, claim 2 was rejected under 35 U.S.C. § 112, second paragraph, as allegedly indefinite, rejection of claims 2-3 was also maintained under 35 U.S.C. § 112, first paragraph, as one skilled in the art clearly would not know how to make or use the invention. In a response to the Second Action, submitted on November 4, 2002 ("response to the Second Action"), Appellants amended claim 2 and added new claims 4-5 to better claim the invention.

A Third and Final Official Action, was issued on January 24, 2003 (the "Final Action"), in which rejection of claims 1-5 was maintained under 35 U.S.C. § 101 as allegedly lacking a patentable utility, claim 5 was rejected under 35 U.S.C. § 112, second paragraph, as being improperly dependent upon itself, rejection of claims 1-5 was also maintained under 35 U.S.C. § 112, first paragraph, as one skilled in the art clearly would not know how to use the invention. In a response to the Final Action, submitted on April 24, 2003 ("response to the Final Action") in which Appellants amended claim 5 to further improve its clarity.

An Advisory Action ("the Advisory Action") was mailed on May 16, 2003, the rejection of claim 5 was withdrawn under 35 U.S.C. § 112, second paragraph, however claims 1-5 remain rejected under 35 U.S.C. § 101 as allegedly lacking a patentable utility, and under 35 U.S.C. § 112, first paragraph, as allegedly one skilled in the art clearly would not know how to use the invention. Therefore, claims 1-5 are the subject of this appeal. A copy of the appealed claims are included below in the Appendix (Section IX).

#### IV. STATUS OF THE AMENDMENTS

For the purposes of Appeal Appellants believe that no additional outstanding amendments exist.

#### V. SUMMARY OF THE INVENTION

The present invention relates to Appellants' discovery and identification of novel human sequences that encode a human metalloprotease, neurolysin. Also disclosed are the tissue expression pattern of these sequences (Page 3, lines 15-21) and naturally occurring polymorphisms that exist within these molecules (page 16, lines 21-28). The specification details a number of uses for the presently claimed sequences, including the detection and diagnosis of human diseases such as, *inter alia*, pain management and cardiac diseases wherein such molecules can act as drug targets (specification at page 1, lines 27-31). Appellants have used the methods described in the specification as filed (page 2 lines 11-32 and page 17, lines 17-23) to construct knockout mice which were used to biologically validate the assertions that the sequences of the present invention have utility as drug targets for human diseases, such as pain management and cardiac diseases (specification at page 1, lines 27-31). Additional uses for the sequences of the present invention include assessing temporal and tissue specific gene expression patterns (specification at page 6, line 24), particularly using a high throughput "chip" format (specification at page 5, line 30 through page 6, line 33), mapping the sequences to a specific region of a human chromosome and identifying protein encoding regions (specification at page 11, line 11-17), determining the genomic structure (specification at page 11, line 5-11), and in diagnostic assays such as forensic analysis, human population biology and paternity determinations (see, for example, the specification from page 11, line 5-17) wherein the sequences of the present invention are particularly useful as the specification identified polymorphisms (page 16, lines 21-28) that can be used in these assays. Thus, Appellants have described novel nucleic and amino acid sequences, their tissue expression pattern and naturally occurring polymorphisms (page 16, lines 21-28) that exist within the sequences of the present invention. The sequences of the present invention encode neurolysin, a protein of known function and Appellants have used methods described in the specification (page 2 lines 11-32 and

page 17, lines 17-23) as filed to biologically validate their assertions that the sequence of the present invention have utility as drug targets for human disease (specification at page 1, lines 27-31) among other utilities.

## **VI. ISSUES ON APPEAL**

1. Do claims 1-5 lack a patentable utility?
2. Are claims 1-5 unusable by a skilled artisan due to a lack of patentable utility?

## **VII. GROUPING OF THE CLAIMS**

For the purposes of the outstanding rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph, the claims will stand or fall together.

## **VIII. ARGUMENT**

### **A. Do Claims 1-5 Lack a Patentable Utility?**

The Final Action rejected and the Advisory Action maintained the rejection of claims 1-5 under 35 U.S.C. § 101, as allegedly lacking a patentable utility due to not being supported by either a specific and substantial utility or a well-established utility, this rejection is maintained in the Advisory Action.

Appellants have described novel nucleic and amino acid sequences, their tissue expression pattern and naturally occurring polymorphisms (page 16, lines 21-28) that exist within the sequences of the present invention. The sequences of the present invention encode neurolysin, a protein of known function and Appellants have used methods described in the specification (page 2 lines 11-32 and page 17, lines 17-23) as filed to biologically validate their assertions that the sequence of the present invention have utility as drug targets for human disease (specification at page 1, lines 27-31).

First, as set forth in the response to the First Action and the response to the Final Action, Appellants would like to invite the Board's attention to the fact that a sequence that is identical at the amino acid level over the entire length of the described sequence is present in the leading

scientific repository for biological sequence data (GenBank), and has been annotated by third party scientists *wholly unaffiliated with Appellants* as neurolysin (*Homo sapiens*) (GenBank accession number: CAC27329, alignment and GenBank report provided in **Exhibit A**). Therefore, it is clear that the amino acid sequence of SEQ ID NO:2 encodes human neurolysin.

Furthermore, Appellants would like to invite the Board's attention to the fact that GenBank Accession No. AJ300837 (alignment and GenBank report provided in **Exhibit B**) that describes a sequence annotated by others to be the mRNA for human neurolysin has a 99% identify with the nucleic acid sequence of SEQ ID NO: 1 (2113 of a total of 2115 bases present in SEQ ID NO: 1). These sequences have also been defined by third party scientists, wholly unaffiliated with Appellants, as encoding human neurolysin. Given this clear evidence that those skilled in the art have independently accepted the utility described in the present specification, there can be no question that Appellants' asserted utility for the described sequences is "credible." As such, the scientific evidence of identity at both the amino acid and nucleic acid levels clearly establishes that those of skill in the art would recognize the sequences of the present invention as human neurolysin, a protein with well known function. Therefore, Appellants have described a utility in full compliance with the provisions of 35 U.S.C. section 101, and the Examiner's rejection should be overturned.

Clearly evidence supports Appellants' assertions that the sequences of the present invention encode a novel human metalloprotease (specifically neurolysin, metallopeptidase M3) which has a well established utility that is recognized by those of skill in the art. Thus this situation parallels Example 10 of the PTO's Revised Interim Utility Guidelines Training Materials (pages 53-55), which establishes that a rejection under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility, is not proper when there is no reason to doubt the asserted utility of a full length sequence (such as the presently claimed sequence) that has a similarity to a protein having a known function. In the Analysis portion of Example 10 it states that "Based on applicant's disclosure and the results of the PTO search, there is no reason to doubt the assertion that SEQ ID NO:2 encodes a DNA ligase. Further DNA ligases have a well-established use in the molecular biology art based on this class of proteins

ability to ligate DNA. ....Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed..... Thus the conclusion reached from this analysis is that a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph, utility rejection should not be made."

The present case is similar to that presented in Example 10 of the Revised Interim Utility Guidelines Training Materials (pages 53-55). In the present case it is clear that the sequences of the present invention encode novel human metalloprotease (specifically neurolysin, metallopeptidase M3) . Neurolysin has well-established utility. According to the guidelines "Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed...Thus the conclusion reached from this analysis is that a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph, utility rejection should not be made." Thus the rejection of the presently claimed invention under a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph utility rejection should be overruled.

The biological significance and function of neurolysin and neurolysin like metalloproteases are well known to those of skill in the art. Evidence includes the publication cited as being allegedly prior art in the First Action under 35 U.S.C. § 102(b)(Kato, *et al.*, J. Biol Chem. 1997, 272, 15313-15322 (**Exhibit C**), in which their role in intracellular degradation of bioactive oligopeptides is described. Appellants submitted one of the many publications describing the activity of neurolysin, in the response to the Second Action in the form of an abstract of a review published by Shrimpton, Smith and Lew (Endocr. Rev. 2002 Oct; 23(5):647-64 (abstract included as **Exhibit D**) and additional evidence is provided by a publication by Norman, *et al.* (Am J Physiol Heart Circ Physiol. 2003: abstract provided as **Exhibit E**). Furthermore, in a section of the Final Action (at last line of page 3) the Examiner states "That are Chen *et al.* who demonstrated that human protein having the amino acid sequence as set forth in SEQ ID NO:2 is neurolysin and has specific, substantial, credible and well established utility." While Appellants disagree with the Examiner's conclusions regarding priority, because the Appellants' present application predates the disclosure of Chen, *et al.* and as stated in the response, that the only reasonable conclusion that one might draw from the information that the Examiner cited is that Chen *et al.* may have been the first to publish "that human protein having

the amino acid sequence as set forth in SEQ ID NO:2 is neurolysin and has specific, substantial, credible and well established utility." However from the Examiner's statements it would appear clear that neurolysin does indeed have specific, substantial, credible and well established utility.

However, in the Advisory Action the Examiner's position is that (page 7, lines 6-10) that "the fact that the third party scientists cloned the human neurolysin gene and disclosed the amino acid sequence of the enzyme after the Applicant filed the application does not change the fact that the protein of SEQ ID NO:2 encoded by DNA of SEQ ID NO:1 was disclosed by Applicants without asserting its function and utility." The Examiner then contradicts her prior statements "That are Chen *et al.* who demonstrated that human protein having the amino acid sequence as set forth in SEQ ID NO:2 is neurolysin and has specific, substantial, credible and well established utility" by taking the position that "even Chen *et al.* have thus far, not disclosed actual enzymatic activity of the protein set forth by SEQ ID NO:2 that is encoded by SEQ ID NO:1" (Advisory Action at page 7 lines 11-12) and presumably therefore, in spite of the many publications to the contrary, the utility of neurolysin remains to be established to the Examiner's satisfaction.

Included in a previous Action as a reason for the alleged lack of utility of the present invention is the statement that "Applicants themselves did not present any evidence that protein of SEQ ID NO:2 is able to cleavage neurotensin between residues Pro10 and Tyr11, and to bind angiotensin; the assays are easy to perform *in vitro*." However, the ease of potential experimentation is not the issue, this emphasis is misplaced as it has long been established that "there is no statutory requirement for the disclosure of a specific example". *In re Gay*, 135 USPQ 311 (C.C.P.A. 1962). Appellants assertion of the stated utility is legally sufficient and should control the utility analysis unless the Examiner meets the burden of establishing the lack of utility by making evidence of record that conclusively refutes the Appellants asserted utility. In the Advisory Action the Examiner submits evidence that she feels does not support Appellants' assertions, however these submissions do not refute Appellants asserted utility.

Additionally, in the Response to the Final Action, Appellants submitted still further evidence of finding supporting the real world utility of the metalloprotease encoded by the sequences of the present invention as provided by Appellants findings involving the analysis of

transgenic "knockout" mice, in which the function of the gene encoding the sequences of the present invention were disrupted in embryonic stem cells (constructed as described in the specification at least on page 2, lines 11-32 and page 17, lines 17-23). Knockout mice prepared as described in the specification as filed were subject to a medical work-up using an integrated suite of medical diagnostic procedures designed to assess the function of the major organ systems in a mammalian subject. Disruption of the mouse gene of the present invention and thus elimination of the protein it encodes, resulted in an increased anxiety-like response, as manifested as a decreased sum center-to-total distance ratio and decreased center-to-total distance travel during all intervals tested when compared with their wild-type (+/+) littermates. Homozygous (-/-) mice, deficient in the gene and protein of the present invention, also demonstrated decreased paw flinching during phase II of a formalin pain assay, as compared to their wild-type (+/+) littermates, suggesting decreased sensitivity to pain. In Addition, female homozygous mutant mice exhibited a decreased heart rate when compared to gender-matched littermates and the historical mean. This clearly provides evidence that the nucleic acid and protein of the present invention have a biological function and the molecules of the present invention as well as agonists or antagonists directed at them can be used to diagnose and treat anxiety and pain disorders as well as cardiac disease (as stated in the specification at least at page 1, line 29), validated drug targets. Thus clearly the molecules of the present invention also have real world substantial and specific utility as having been identified as biologically validated drug targets using methods and identified for diseases and disorders asserted in the specification as filed.

Appellants invite the board to consider that those of skill in the art would readily recognize that molecules which share amino acid identity would also share function. Thus as SEQ ID NO:2 is 100% identical with a molecule recognized by those of skill in the art as neurolysin, SEQ ID NO:2 would also be recognized to have all the function, specific, substantial, credible and well established utility of neurolysin. The Advisory Action states (the last line of page 7-line 2 of page 8) that "Furthermore, the availability of post-filing date evidence of utility is irrelevant in the instant situation, where Applicants specification lacks any assertion of this utility." This position contradicts the position presented in Example 10 of the Revised

Interim Utility Guidelines Training Materials (pages 53-55), "Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed..."

In addition to contradicting Example 10 of the Guidelines, when the Advisory Action states (page 4 line 3-6) that Appellants describe only a protein sharing sequence similarity with mammalian neurolysin there is no assertion of function. Appellants disagree, as clearly those of skill in the art would readily recognize that functional properties are inherent to structural identity, identical structure begets identical function and thus, logically, identical utility.

Furthermore, Appellants have made several presentations which the Examiner has found non-persuasive as described in the Advisory Action. The Examiner does not appear to accept that Appellants assertion that the sequences of the present invention encode a novel human metalloprotease specifically metallopeptidase M3, neurolysin. In fact the Examiner does not appear to believe that such assertions were even made in the specification. Appellants respectfully submit to the Board that such assertions were made, among other locations, in the title of the original application "Novel Human Metalloprotease and Polynucleotides Encoding the Same"; in the background information section of the specification, Section 2, where Appellants describe the activity of metalloproteases including neurolysin; and in the position taken by the Appellants throughout the specification and prosecution wherein the Appellants have identified the structural similarities between mammalian neurolysin proteins and the sequences of the present invention. Appellants have provided several pieces of evidence that those of skill in the art would find Appellants assertions credible. That evidence clearly shows that those of skill in the art, in no way affiliated with Appellants, when faced with the same information, would and did identify the sequences of the present invention as a metalloprotease, neurolysin and that those of skill in the art would readily recognize that functional properties are inherent to structural identity, identical structure begets identical function and thus, logically, identical utility.

Thus, the skilled artisan would readily appreciate the utilities asserted by Appellants' regarding the role of the proteins encoded by sequences of the present invention, including those associated with diseases that have been linked to the novel human metalloprotease, neurolysin.

Therefore, the present utility rejection must fail. According to the Examination Guidelines for the Utility Requirement, if the applicant has asserted that the claimed invention is useful for any particular purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, the Examiner should not impose a rejection based on lack of utility (66 Federal Register 1098, January 5, 2001).

The Advisory Action maintains that Appellants' assertions regarding the use of the presently claimed polynucleotides on DNA gene chips, based on the position that such a use would allegedly be generic. Further, these Actions seem to be requiring Appellants to identify the biological role of the nucleic acid or function of the protein encoded by the presently claimed polynucleotides before the present sequences can be used in gene chip applications that meet the requirements of § 101. Appellants respectfully point out that knowledge of the exact function or role of the presently claimed sequence is not required to track expression patterns using a DNA chip. As set forth in Appellants First Response, given the widespread utility of such "gene chip" methods using *public domain* gene sequence information, there can be little doubt that the use of the presently described *novel* sequences would have great utility in such DNA chip applications.

However, clearly given the extensive utility described above for the molecules encoded by the sequences of the present invention and evidence that the claimed sequences provide a specific marker of the gene encoding neurolysin and provide a unique identifier of the corresponding gene in the human genome. Such specific markers are targets for discovering drugs that are associated with human disorders and diseases such as, *inter alia*, pain management and cardiac disease (specification page 1, line 29). Thus, those skilled in the art would instantly recognize that the present nucleotide sequence would be an ideal, novel candidate for assessing gene expression using, for example, DNA chips, as the specification details at least on page 5, line 30 through page 6, line 33. Such "DNA chips" clearly have utility, as evidenced by hundreds of issued U.S. Patents, exemplified by U.S. Patent Nos. 5,445,934 (**Exhibit F**), 5,556,752 (**Exhibit G**), 5,744,305 (**Exhibit H**), as well as more recently issued U.S. Patent Nos. 5,837,832 (**Exhibit I**), 6,156,501 (**Exhibit J**) and 6,261,776 (**Exhibit K**).

The Board is further requested to consider that, given the huge expense of the drug discovery process, even negative information has great "real world" practical utility. Knowing

that a given gene is not expressed in medically relevant tissue provides an informative finding of great value to industry by allowing for the more efficient deployment of expensive drug discovery resources. Such practical considerations are equally applicable to the scientific community in general, in that time and resources are not wasted chasing what are essentially scientific dead-ends (from the perspective of medical relevance). Clearly, compositions that enhance the utility of such DNA gene chips, such as the presently claimed sequences encoding neurolysin which Appellants have shown are biologically validated drug targets for pain and cardiac disorder, among others, must in themselves be useful. Moreover, the presently described protein (neurolysin) provides uniquely specific sequence resources for identifying and quantifying full length transcripts that were encoded by the corresponding human genomic locus. Accordingly, there can be no question that the described sequences provide an exquisitely specific utility for analyzing gene expression.

The utility of the sequences of the present invention are further enhanced by the description in the specification of tissues expressing the sequences of the present invention (Page 3, lines 15-21) and the description of polymorphisms (page 16, lines 21-28). These teachings along with the above evidence that the molecules of the present invention encode a protein of known function and that Appellants have used methods described in the specification as filed to biologically validate their assertions that the sequence of the present invention have utility as drug targets for human disease, clearly demonstrate outstanding utility of the sequences in DNA chip expression analysis.

Still further, as only a small percentage of the genome (2-4%) actually encodes exons, which in turn encode amino acid sequences. Thus, not all human genomic DNA sequences are useful in such gene chip applications. This further discounts the Examiner's position that such uses are "generic". The present claims clearly meet the requirements of 35 U.S.C. § 101. It has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such statement. *In re Langer*, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971).

Additional evidence of the "real world" substantial utility of the present invention is further provided by the fact that there is an entire industry based on the use of gene sequences or fragments thereof in a gene chip format. Perhaps the most notable gene chip company is Affymetrix. However, there are many companies which have, at one time or another, concentrated on the use of gene sequences or fragments, in gene chip and non-gene chip formats, for example: Gene Logic, ABI-Perkin-Elmer, HySeq and Incyte. In addition, one such company, Rosetta Inpharmatics, was viewed to have such "real world" value that it was acquired by large pharmaceutical company, Merck & Co., for substantial sums of money (net equity value of the transaction was \$620 million). The "real world" substantial industrial utility of gene sequences or fragments would, therefore, appear to be widespread and well established. Clearly, persons of skill in the art, as well as venture capitalists and investors, readily recognize the utility, both scientific and commercial, of genomic data in general, and specifically human genomic data. Billions of dollars have been invested in the human genome project, resulting in useful genomic data (see, *e.g.*, Venter *et al.*, 2001, Science 291:1304; **Exhibit L**). The results have been a stunning success as the utility of human genomic data has been widely recognized as a great gift to humanity (see, *e.g.*, Jasny and Kennedy, 2001, Science 291:1153; **Exhibit M**). Clearly, the usefulness of human genomic data, such as the presently claimed nucleic acid molecules, is substantial and credible (worthy of billions of dollars and the creation of numerous companies focused on such information) and well-established (the utility of human genomic information has been clearly understood for many years).

As a still further example of utility is the use of the present sequences in such diagnostic assays (at least at page 11, lines 5-17) as those associated with identification of paternity and forensic analysis, among others. The sequences of the present invention have particular utility as the application as filed identified several polymorphisms (page 16, lines 21-28). This is also not a case of a potential utility. Appellants respectfully submit that even in the worst case scenario, the described polymorphisms are each useful to distinguish 50% of the population (in other words, the marker being present in half of the population) and that the ability of a polymorphic marker to distinguish at least 50% of the population is an inherent feature of any polymorphic marker, and this feature is well understood by those of skill in the art. Appellants note that as a

matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988). Appellants support for Appellants' assertion of utility is provided by the fact that the skilled artisan would readily recognize and easily believe that the presently described polymorphic markers could be useful in forensic analysis. The fact that forensic biologists use polymorphic markers such as those described by Appellants every day provides more than ample support for the assertion that forensic biologists would also be able to use the specific polymorphic markers described by Appellants in the same fashion. Therefore, again it is clear that the sequences of the present invention have utility.

Given the physiologic activity and importance of neurolysin as known to those of skill in the art, those of skill in the art would readily appreciate the importance of tracking the expression of the genes encoding the described proteins, particularly due to the established role of neurolysin in human disorders such as such as pain management and cardiac diseases (specification at page 1, lines 27-31). The use of the claimed polypeptide in an array for screening purposes Appellants respectfully point out that nucleic acid sequences have the greatest specific utility in gene chip applications once the role of the sequence has been identified, as have tissues of interest, as in the present case. Once the role of the particular nucleic acid is known, the level of gene expression has and even greater significance. By identifying the physiological activity role of the claimed sequence, the claimed sequence has a far greater utility in gene chip applications than just any random piece of DNA. Appellants respectfully submit that specific utility, which is the proper standard for utility under 35 U.S.C. § 101, is distinct from the requirement for a unique utility, which is clearly an improper standard. As clearly stated by the Federal Circuit in *Carl Zeiss Stiftung v. Renishaw PLC*, 20 USPQ2d 1101 (Fed. Cir. 1991; "Carl Zeiss"):

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding a lack of utility." *Envirotech Corp. v. Al George, Inc.*, 221 USPQ 473, 480 (Fed. Cir. 1984)

Therefore, just because other nucleic acid sequences find utility in gene chip applications does not mean that the use of Appellants' sequence in gene chip applications is not a specific utility. Furthermore, the requirement for a unique utility is clearly not the standard adopted by the Patent

and Trademark Office. If every invention were required to have a unique utility, the Patent and Trademark Office would no longer be issuing patents on batteries, automobile tires, golf balls, golf clubs, and treatments for a variety of human diseases, such as cancer and bacterial or viral infections, just to name a few particular examples, because examples of each of these have already been described and patented. All batteries have the exact same utility - specifically, to provide power. All automobile tires have the exact same utility - specifically, for use on automobiles. All golf balls and golf clubs have the exact same utility - specifically, use in the game of golf. All cancer treatments have the exact same utility - specifically, to treat cancer. All anti-infectious agents have the exact same broader utility - specifically, to treat infections. However, only the briefest perusal of virtually any issue of the Official Gazette provides numerous examples of patents being granted on each of the above compositions every week. Furthermore, if a composition needed to be unique to be patented, the entire class and subclass system would be an effort in futility, as the class and subclass system serves solely to group such common inventions, which would not be required if each invention needed to have a unique utility. Thus, the present sequence clearly meets the requirements of 35 U.S.C. § 101.

Further evidence of utility of the presently claimed polynucleotide, although only one is needed to meet the requirements of 35 U.S.C. § 101 (*Raytheon v. Roper*, 220 USPQ 592 (Fed. Cir. 1983); *In re Gottlieb*, 140 USPQ 665 (CCPA 1964); *In re Malachowski*, 189 USPQ 432 (CCPA 1976); *Hoffman v. Klaus*, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988)), is the specific utility the present nucleotide sequence has in determining the genomic structure of the corresponding human chromosome (specification at page 11, lines 5-11), for example mapping the protein encoding regions as described in the specification (page 11, lines 11-17) and evidenced below. Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of the human chromosome containing the gene encoding the given polynucleotide, a utility not shared by virtually any other nucleic acid sequence. In fact, it is this specificity that makes this particular sequence so useful. Early gene mapping techniques relied on methods such as Giemsa staining to identify regions of chromosomes. However, such techniques produced genetic maps with a resolution of only 5 to 10 megabases, far too low to be of much help in identifying specific genes involved in disease. The skilled artisan readily

appreciates the significant benefit afforded by markers that map a specific locus of the human genome, such as the present nucleic acid sequence.

Only a minor percentage of the genome actually encodes exons, which in turn encode amino acid sequences. The presently claimed polynucleotide sequence provides biologically validated empirical data (e.g., showing which sequences are transcribed, spliced, and polyadenylated) that *specifically* defines that portion of the corresponding genomic locus that actually encodes exon sequence. Equally significant is that the claimed polynucleotide sequence defines how the encoded exons are actually spliced together to produce an active transcript (i.e., the described sequences are useful for functionally defining exon splice-junctions). The Appellants respectfully submit that the practical scientific value of expressed, spliced, and polyadenylated mRNA sequences is readily apparent to those skilled in the relevant biological and biochemical arts. For further evidence supporting the Appellants' position, the Board is requested to review, for example, section 3 of Venter *et al.* (*supra* at pp. 1317-1321, including Fig. 11 at pp.1324-1325), which demonstrates the significance of expressed sequence information in the structural analysis of genomic data. The presently claimed polynucleotide sequence defines a biologically validated sequence that provides a unique and specific resource for mapping the genome essentially as described in the Venter *et al.* article.

As still further evidence supporting Appellants assertions of the specific utility of the sequences of the present invention in localizing the specific region of the human chromosome and identification of functionally active intron/exon splice junctions is the information provided in **Exhibit N**. This is the result of a blast analysis using SEQ ID NO:1 of the present invention when compared to the identified human genomic sequence. This result indicates that the sequence of the present invention is encoded by 13 exons spread non-contiguously along a region of human chromosome 5, which are contained within partially overlapping BAC clones, AC016643.6 and AC008958.7. Thus clearly one would not simply be able to identify the 13 protein encoding exons that make up the sequence of the present intention from within the large genomic sequence. Nor, would one be able to map the protein encoding regions identified specifically by the sequences of the present invention without knowing exactly what those specific sequences were.

In addition to the previously submitted Exhibits, demonstrating that the sequences of the present invention encode a metalloprotease, neurolysin. Genetic mapping of the sequences of the present invention maps to human chromosome 5, at the very same region as that to which neurolysin (NLN) has been mapped. Therefore, in addition to the clear sequence homology between molecules annotated as neurolysin, a metalloprotease, and the sequences of the present invention which have been evidenced. The sequences of the present invention and neurolysin (NLN), map to the same genetic locus, 5q12.2.

The question of utility is a straightforward one. As set forth by the Federal Circuit, "(t)he threshold of utility is not high: An invention is 'useful' under section 101 if it is capable of providing some identifiable benefit." *Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)). Additionally, the Federal Circuit has stated that "(t)o violate § 101 the claimed device must be totally incapable of achieving a useful result." *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992), emphasis added. *Cross v. Iizuka* (224 USPQ 739 (Fed. Cir. 1985); "Cross") states "any utility of the claimed compounds is sufficient to satisfy 35 U.S.C. § 101". *Cross* at 748, emphasis added. Indeed, the Federal Circuit recently emphatically confirmed that "anything under the sun that is made by man" is patentable (*State Street Bank & Trust Co. v. Signature Financial Group Inc.*, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998), citing the U.S. Supreme Court's decision in *Diamond vs. Chakrabarty*, 206 USPQ 193 (S.Ct. 1980)).

The legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be credible or believable. According to the Examination Guidelines for the Utility Requirement, if the applicant has asserted that the claimed invention is useful for any particular purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, the Examiner should not impose a rejection based on lack of utility (66 Federal Register 1098, January 5, 2001).

In *In re Brana*, (34 USPQ2d 1436 (Fed. Cir. 1995), "Brana"), the Federal Circuit admonished the P.T.O. for confusing "the requirements under the law for obtaining a patent with

the requirements for obtaining government approval to market a particular drug for human consumption". *Brana* at 1442. The Federal Circuit went on to state:

At issue in this case is an important question of the legal constraints on patent office examination practice and policy. The question is, with regard to pharmaceutical inventions, what must the applicant provide regarding the practical utility or usefulness of the invention for which patent protection is sought. This is not a new issue; it is one which we would have thought had been settled by case law years ago.

*Brana* at 1439, emphasis added. The choice of the phrase "utility or usefulness" in the foregoing quotation is highly pertinent. The Federal Circuit is evidently using "utility" to refer to rejections under 35 U.S.C. § 101, and is using "usefulness" to refer to rejections under 35 U.S.C. § 112, first paragraph. This is made evident in the continuing text in *Brana*, which explains the correlation between 35 U.S.C. §§ 101 and 112, first paragraph. The Federal Circuit concluded:

FDA approval, however, is not a prerequisite for finding a compound useful within the meaning of the patent laws. Usefulness in patent law, and in particular in the context of pharmaceutical inventions, necessarily includes the expectation of further research and development. The stage at which an invention in this field becomes useful is well before it is ready to be administered to humans. Were we to require Phase II testing in order to prove utility, the associated costs would prevent many companies from obtaining patent protection on promising new inventions, thereby eliminating an incentive to pursue, through research and development, potential cures in many crucial areas such as the treatment of cancer.

*Brana* at 1442-1443, citations omitted. In assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation". *In re Angstadt and Griffin*, 190 USPQ 214 (C.C.P.A. 1976). The need for some experimentation does not render the claimed invention unpatentable. Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin*, *supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 18 USPQ2d 1016 (Fed. Cir. 1991). As a matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988).

Finally, with regards to the issue of due process, while Appellants are well aware of the new Utility Guidelines set forth by the USPTO, Appellants respectfully point out that the current rules and regulations regarding the examination of patent applications is and always has been the patent laws as set forth in 35 U.S.C. and the patent rules as set forth in 37 C.F.R., not the Manual of Patent Examination Procedure or particular guidelines for patent examination set forth by the USPTO. Furthermore, it is the job of the judiciary, not the USPTO, to interpret these laws and rules. Appellants are unaware of any significant recent changes in either 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit that is in keeping with the new Utility Guidelines set forth by the USPTO. This is underscored by numerous patents that have been issued over the years that claim nucleic acid fragments that do not comply with the new Utility Guidelines. As examples of such issued U.S. Patents, the Board is invited to review U.S. Patent Nos. 5,817,479 (**Exhibit O**), 5,654,173 (**Exhibit P**), and 5,552,281 (**Exhibit Q**; each of which claims short polynucleotides), and recently issued U.S. Patent No. 6,340,583 (**Exhibit R**; which includes no working examples), none of which contain examples of the “real-world” utilities that the Examiner seems to be requiring. As issued U.S. Patents are presumed to meet all of the requirements for patentability, including 35 U.S.C. §§ 101 and 112, first paragraph (see Section VIII(B), below), Appellants submit that the present polynucleotides must also meet the requirements of 35 U.S.C. § 101. While Appellants agree that each application is examined on its own merits, Appellants are unaware of any changes to 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit, since the issuance of these patents that render the subject matter claimed in these patents, which is similar to the subject matter in question in the present application, as suddenly non-statutory or failing to meet the requirements of 35 U.S.C. § 101. Given the rapid pace of development in the biotechnology arts, it is difficult for the Appellants to understand how an invention fully disclosed and free of prior art at the time the present application was filed, could somehow retain *less* utility and be *less* enabled than inventions in the cited issued U.S. patents (which were filed during a time when the level of skill in the art was clearly lower). Simply put, Appellants invention is *more* enabled and retains *at least as much* utility as the inventions described in the claims of the U.S. patents of record. Thus, holding Appellants to a different

standard of utility would be arbitrary and capricious, and, like other clear violations of due process, cannot stand.

Thus in summary, Appellants have described novel nucleic and amino acid sequences, their tissue expression pattern and naturally occurring polymorphisms that exist within these molecules. Furthermore, the sequences of the present invention encode the human metalloprotease, neurolysin, a protein of well recognized function. In addition, Appellants have used methods described in the specification as filed to biologically validate their assertions that the sequences of the present invention have utility as drug targets for human disease. Thus, the conclusion reached from this analysis is that a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph, utility rejection should not have been made. Therefore, Appellants respectfully submit that the rejection of the presently claimed invention under a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph utility rejection should be overruled.

#### **B. Are Claims 1-5 Unusable Due to a Lack of Patentable Utility?**

The Final Action next rejects claims 1-5 under 35 U.S.C. § 112, first paragraph, since allegedly one skilled in the art would not know how to use the invention, as the invention allegedly is not supported by either a clear asserted utility or a well-established utility.

The arguments detailed above in **Section VIII(A)** concerning the utility of the presently claimed sequences are incorporated herein by reference. As the Federal Circuit and its predecessor have determined that the utility requirement of Section 101 and the how to use requirement of Section 112, first paragraph, have the same basis, specifically the disclosure of a credible utility (*In re Brana, supra*; *In re Jolles*, 628 F.2d 1322, 1326 n.11, 206 USPQ 885, 889 n.11 (CCPA 1980); *In re Fouche*, 439 F.2d 1237, 1243, 169 USPQ 429, 434 (CCPA 1971)), Appellants submit that as claims 1-5 have been shown to have “a specific, substantial, and credible utility”, as detailed in **Section VIII(A)** above, the present rejection of claims 1-5 under 35 U.S.C. § 112, first paragraph, cannot stand.

Appellants therefore submit that the rejection of claims 1-5 under 35 U.S.C. § 112, first paragraph, must be overruled.

## IX. APPENDIX

The claims involved in this appeal are as follows:

1. An isolated nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO:1.
2. An isolated nucleic acid molecule comprising a nucleotide sequence that:
  - (a) encodes the amino acid sequence shown in SEQ ID NO: 2; and
  - (b) hybridizes under highly stringent conditions including washing in 0.1xSSC/0.1% SDS at 68°C to the nucleotide sequence of SEQ ID NO: 1 or the complement thereof.
3. An isolated nucleic acid molecule comprising a nucleotide sequence encoding the amino acid sequence shown in SEQ ID NO:2.
4. An expression vector comprising a nucleic acid sequence of Claim 3.
5. A cell comprising the expression vector of Claim 4.

## X. CONCLUSION

Appellants respectfully submit that, in light of the foregoing arguments, the Final Action's conclusion that claims 1-5 lack a patentable utility and are unusable by the skilled artisan due to a lack of patentable utility is unwarranted. It is therefore requested that the Board overturn the Final Action's rejections.

Respectfully submitted,

September 29, 2003

Date

  
Lance K. Ishimoto

Reg. No. 41,866

Agent For Appellants

LEXICON GENETICS INCORPORATED  
8800 Technology Forest Place  
The Woodlands, TX 77381  
(281) 863-3399

**Customer # 24231**

FASTA searches a protein or DNA sequence data bank  
version 3.3t05 March 30, 2000

Please cite:

W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

```
/tmp/fastaFAA4GaWRi: 704 aa
>SEQ ID NO:2
vs /tmp/fastaGAA5GaWRi library
searching /tmp/fastaGAA5GaWRi library
```

704 residues in 1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2  
join: 38, opt: 26, gap-pen: -12/ -2, width: 16

Scan time: 0.033

The best scores are: opt
gi|12583612|emb|CAC27329.1| neurolysin [Homo sapi ( 704) 4666

>>gi|12583612|emb|CAC27329.1| neurolysin [Homo sapiens]
(704 aa)

initn: 4666 init1: 4666 opt: 4666

Smith-Waterman score: 4666; 100.000% identity in 704 aa overlap (1-704:1-704)

	10	20	30	40	50	60	
SEQ	MIARCLLAVSLRRVGGSRILLRMTLGREVMSPLQAMSSYTVAGRNVLRWDLSPEQIKTR	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	MIARCLLAVSLRRVGGSRILLRMTLGREVMSPLQAMSSYTVAGRNVLRWDLSPEQIKTR	10	20	30	40	50	60
	70	80	90	100	110	120	
SEQ	TEELIVQTKQVYDAVGMLGIEEVTYENCLQALADVEVKYIVERTMLDFPQHVSSDKEVRA	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	TEELIVQTKQVYDAVGMLGIEEVTYENCLQALADVEVKYIVERTMLDFPQHVSSDKEVRA	70	80	90	100	110	120
	130	140	150	160	170	180	
SEQ	ASTEADKRLSRFDIEMSMRGDIFERIVHLQETCDLGKIKPEARRYLEKSIKMGKRNLHL	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	ASTEADKRLSRFDIEMSMRGDIFERIVHLQETCDLGKIKPEARRYLEKSIKMGKRNLHL	130	140	150	160	170	180
	190	200	210	220	230	240	
SEQ	PEQVQNEIKSMKKRMSEL CIDFNKNLNE DDTFLVFSKAELGALPDDFIDSLEKTDDDKYK	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	PEQVQNEIKSMKKRMSEL CIDFNKNLNE DDTFLVFSKAELGALPDDFIDSLEKTDDDKYK	190	200	210	220	230	240
	250	260	270	280	290	300	
SEQ	ITLKYPHYFPVMKKCCIPETRRRMEMAFNTRCKEENTIILQQQLLPLRTKVAKLLGYSTHA	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	ITLKYPHYFPVMKKCCIPETRRRMEMAFNTRCKEENTIILQQQLLPLRTKVAKLLGYSTHA	250	260	270	280	290	300
	310	320	330	340	350	360	
SEQ	DFVLEMNTAKSTSRVTAFLDDLSQKLKPLGEAERE FILNLKKKECKDRGF EYDGKINAWD	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	:::::::::::::::::::	
gi 125	DFVLEMNTAKSTSRVTAFLDDLSQKLKPLGEAERE FILNLKKKECKDRGF EYDGKINAWD	310	320	330	340	350	360
	370	380	390	400	410	420	

SEQ	LYYYMTQTEELKYSIDQEFLKEYFPIEVVTEGLLN	TYQELLGLSFEQMTDAH	VWNKSVTL			
gi 125	LYYYMTQTEELKYSIDQEFLKEYFPIEVVTEGLLN	TYQELLGLSFEQMTDAH	VWNKSVTL			
	370	380	390	400	410	420
SEQ	430	440	450	460	470	480
gi 125	YTVKDKATGEVLGQFYLDLYPREGKYNHAA	CFG	QPG	CLL	PDG	SRM
	430	440	450	460	470	480
SEQ	490	500	510	520	530	540
gi 125	490	500	510	520	530	540
SEQ	550	560	570	580	590	600
gi 125	550	560	570	580	590	600
SEQ	610	620	630	640	650	660
gi 125	610	620	630	640	650	660
SEQ	670	680	690	700		
gi 125	670	680	690	700		

704 residues in 1 query sequences

704 residues in 1 library sequences

Scomplib [version 3.3t05 March 30, 2000]

start: Fri Nov 1 15:59:40 2002 done: Fri Nov 1 15:59:41 2002

Scan time: 0.033 Display time: 0.700

Function used was FASTA



ORIGIN

1 ggcacgaggg tggccctctgc ggctaggccg gctcagact cccggcgcc gaggcgctgc  
61 cggccgcctc gccgcggccac gccgaaggac cacgcgcccgg cccgcggccag cctctcagcg  
121 ctccccatgtat cggccgggtgc cttttggctg tgcgaagcct ccgcagagtt ggtggttcca  
181 ggatTTTact cagaatgacg ttaggaagag aagtatgtc tcctctttag gcaatgtctt  
241 cctatactgt ggctggcaga aatgtttaa gatggatct ttacccagag caaattaaaa  
301 caagaactga ggagctcatt gtgcagacca aacaggtgt a cgtgctgtt ggaatgctcg  
361 gtattgagga agtaacttac gagaactgtc tgccggact ggcagatgt a gaagtaaagt  
421 atatagtgga aaggaccatg ctagacttcc cccagcatgt atccctctgac aaagaagtac  
481 gaggcagcaag tacagaagca gacaaaagac tttctcggtt tgatatttag atgagcatga  
541 gaggagatat atttgagaga attgttcatt tacaggaaac ctgtgatctg gggaaagataa  
601 aacctgaggg cagacgatac ttggaaaagt caattaaaaat ggggaaaaga aatgggctcc  
661 atcttcctgaa cacaatgtac aatgaaatca aatcaatgaa gaaaagaatg agtgagctat  
721 gtattgattt taacaaaaac ctcataatgagg atgataaccc ctttgcgtt tccaaggctg  
781 aacttgggtgc tcttcctgat gatttcattt acagttttaga aaagacagat gatgacaagt  
841 ataaaaattac cttaaaaatcc ccaacttatt tccctgtcat gaagaaatgt tgatccctg  
901 aaaccagaag aaggatggaa atggcttttata atacaagggtg caaagaggaa aacaccataa  
961 ttttgcagca gctactccca ctgcgaacca aggtggccaa actactcggt tatagcacac  
1021 atgctgactt cgtccttggaa atgaacactg caaagagcac aagccgcgtt acagccttcc  
1081 tagatgattt aagccagaag ttaaaaccct tgggtgaagc agaacgagag tttatTTGA  
1141 atttgaagaa aaaggaatgc aaagacaggg gtttgaata tgatggaaa atcaatgcct  
1201 gggatctata ttactacatg actcagacac aggaactcaa gtattccata gaccaagagt  
1261 tcctcaagga atacttccca attgaggtgg tcaactgaagg cttgctgaac acctaccagg  
1321 agttgttggg actttcattt gaacaaatgtc cagatgtca tgtttggAAC aagagtgtta  
1381 cactttatac tgtgaaggat aaagctacag gagaagtatt gggacagttc tatttggacc  
1441 tctatccaag ggaaggaaaa tacaatcatg cggccgtt cggctccag cctggctgcc  
1501 ttctgcctga tggaaagccgg atgatggcag tggctgcct cgtggtaac ttctcacagc  
1561 cagtggcagg tgcgtccctt ctcctgagac acgacgaggt gaggacttac tttcatgagt  
1621 ttggtcacgt gatgcacatg atttgcac agactgattt tgacacgattt agcggAACAA  
1681 atgtggaaac tgactttgtt gaggtgcattt cgcaaatgtc tgaaaattgg gtgtgggacg  
1741 tcgattccctt ccgaagatttgc taaaacattt ataaagatgg aagccctattt gcaagacgatc  
1801 tgcttgaaaa acttgcgttgc tctaggctgg tcaacacagg tcttctgacc ctgcgcaga  
1861 ttgttttgag caaagttgtt cagtcttttcc ataccaacac atcgctggat gctgcaagtgc  
1921 aatatgccaa atactgtca gaaatattttttag gagttgcagc tactccaggc acaaataatgc  
1981 cagctaccc ttggacatttgc cggggggat acgatggccaa atattatggat tatctttggaa  
2041 gtgaagtattt ttccatggat atgttttaca gctgtttttaaaaaggaaatgggataatgtac  
2101 cagaggttgg aatgaaatac agaaacctaa tcctgaaacc tggggatct ctggacggca  
2161 tggacatgtt ccacaatttgc ttggaaacgtt agccaaacca aaaagcggtt ctaatgatgt  
2221 gaggcctgca tgctcgttgc actggggatc ttggtagcc gtccatgttgc ggaggacaag  
2281 tcgacatcac catgttttac tggcctggaa actgaaggatc gttttgcacat tgaaaatttt  
2341 gatttctattt gacatcctt ttgtttctttaaaaat tataaagatg taaatggat  
2401 tataaataactt gtgacctaag aaaagaccca ctagaaagttt atgttactt aaaaatttt  
2461 aaaactggat ttgatttttgc ttatggaaat tttcatatgtc atgttaacttgc atttttact  
2521 attataatctt agataatatg atataagagg gctaaatgtt tttaaatgtc atcatatata  
2581 tgatataattt tgatccttgc ttgtatcttgc agttttgtac ttgggatcc tggactgtata  
2641 aatgaatcatc cacatttttgc ttggaaatattt tttcttgcacttgcgttca actttgtatcc  
2701 ttgttctccc aggaagggtt gacccctccat ttgcctgttgc acctcaaggc cagggaaata  
2761 tgcctcgttgc atgcattttat ctttgcatttgc caggccgttgc gattccaaac ttctgcac  
2821 acttaaattt cgttccatccat ttgcattttgc tctaaatgttgc agtcaaaagat  
2881 tatcaaaaaaa

//

Revised: October 24, 2001.

[Disclaimer](#) | [Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)

# Targeting of Endopeptidase 24.16 to Different Subcellular Compartments by Alternative Promoter Usage\*

(Received for publication, February 28, 1997)

Akira Kato†, Naoaki Sugiura§, Yohko Saruta, Takehiko Hosoiri, Hiroshi Yasue¶, and Shigehisa Hirose||

From the Department of Biological Sciences, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226, Japan, the §Department of Biochemistry I, National Defense Medical College, Tokorosawa, Saitama 359, Japan, and the ¶Animal Genome Research Group, National Institute of Animal Industry, Inashiki-gun, Ibaraki 305, Japan

Endopeptidase 24.16 or mitochondrial oligopeptidase, abbreviated here as EP 24.16 (MOP), is a thiol- and metal-dependent oligopeptidase that is found in multiple intracellular compartments in mammalian cells. From an analysis of the corresponding gene, we found that the distribution of the enzyme to appropriate subcellular locations is achieved by the use of alternative sites for the initiation of transcription. The pig EP 24.16 (MOP) gene spans over 100 kilobases and is organized into 16 exons. The core protein sequence is encoded by exons 5–16 which match perfectly with exons 2–13 of the gene for endopeptidase 24.15, another member of the thimet oligopeptidase family. These two sets of 11 exons share the same splice sites, suggesting a common ancestor. Multiple species of mRNA for EP 24.16 (MOP) were detected by the 5'-rapid amplification of cDNA ends and they were shown to have been generated from a single gene by alternative choices of sites for the initiation of transcription and splicing. Two types of transcript were prepared, corresponding to transcription from distal and proximal sites. Their expression *in vitro* in COS-1 cells indicated that they encoded two isoforms (long and short) which differed only at their amino termini: the long form contained a cleavable mitochondrial targeting sequence and was directed to mitochondria; the short form, lacking such a signal sequence, remained in the cytosol. The complex structure of the EP 24.16 (MOP) gene thus allows, by alternative promoter usage, a fine transcriptional regulation of coordinate expression, in the different subcellular compartments, of the two isoforms arising from a single gene.

Metalloendopeptidases form a large family of peptidases that have a His-Glu-X-X-His (HEXXH) zinc-binding motif and preferentially cleave short substrates. For example, endopeptidase 24.15 (EP<sup>1</sup> 24.15), a member of this family, acts on peptides of 6–18 amino acid residues and exhibits no or only very weak

proteolytic activity against proteins (1–3). Among the members of this family, thimet oligopeptidase (TOP or EP 24.15)<sup>1</sup> and oligopeptidase M (MOP or EP 24.16) are unique in their sensitivities to thiol reagents and they constitute a subfamily, the thimet (thiol- and metal-dependent) oligopeptidase subfamily. Recent molecular cloning revealed the presence of a cysteine residue unique to members of this subfamily near position 483. This residue is absent from the other members that exhibit no thiol dependence (4, 5). In addition to the members of this family of mammalian origin, certain oligopeptidases of microbial origin that belong to this family have also been identified, including oligopeptidase A (OpdA) and dipeptidyl carboxypeptidase (Dcp) of *Escherichia coli* and *Salmonella typhimurium* (6), peptidase F of *Lactococcus lactis* (7), mitochondrial intermediate peptidase of rat and yeast (8, 9), and saccharolysin (YCL57w or proteinase yscD) of yeast (10). This report deals with the two best characterized mammalian enzymes, namely, EP 24.15 (TOP) and EP 24.16 (MOP), which are members of the thimet oligopeptidase family. This family has also been called the M3 family of metalloendopeptidases in the classification of Rawlings and Barrett (11, 12).

EP 24.15 (TOP) was first identified as a collagenase-like peptidase or Pz-peptidase in experiments with the Pz-peptide that was originally designed by Wünsch and Heidrich (13) as a substrate for collagenase. Although the Pz-peptide was a good substrate for clostridial collagenase, it turned out not to be a substrate for avian and mammalian collagenases (14). The Pz-peptide hydrolyzing activities found in avian and mammalian tissues have, therefore, been designated collagenase-like peptidases or simply Pz-peptidases. Independent studies on the metabolism of brain peptides led to the discovery of two enzymes: one was described by Camargo *et al.* (15) in 1972 and was named neutral endopeptidase and, later, endo-oligopeptidase A; and the other, first described by Orlowski *et al.* (16) in 1983, was initially named soluble metalloendopeptidase and subsequently endopeptidase 24.15. All these enzymes turned out to be the same and are now known as thimet oligopeptidase (17). In this report we use the abbreviated designation EP 24.15 (TOP). cDNA sequences for the mammalian enzyme are now available for the rat (4, 18, 19), pig (20), and human (21).

EP 24.16 (MOP) was also discovered independently in several different laboratories. 1) Heidrich *et al.* (22) demonstrated a Pz-peptide hydrolyzing activity in a mitochondrial fraction of rat liver, which was later shown to be distinct from EP 24.15 (TOP) by both biochemical characterization (23) and partial

\* This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (Intracellular Proteolysis) from the Ministry of Education, Science, Sports and Culture of Japan and by research grants from SRF, the Cosmetology Research Foundation, and the Life Insurance Association of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AB000170–AB000175, AB000411–AB000425, and AB000426–AB000438.

† Supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science.

‡ To whom correspondence should be addressed: Dept. of Biological Sciences, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226, Japan. Tel: 81-45-924-5726; Fax: 81-45-924-5824.

<sup>1</sup> The abbreviations used are: TOP, thimet oligopeptidase; EP, endopeptidase; MOP, oligopeptidase M; PAGE, polyacrylamide gel electrophoresis; bp, base pair(s); SINE(s), short interspersed repetitive element(s); RACE, rapid amplification of cDNA ends; PIPES, 1,4-piperazinediethanesulfonic acid; nt, nucleotide(s).

amino acid sequencing of the purified enzyme; it was named oligopeptidase M (24). 2) We (25) and Kiron and Soffer (26) identified a soluble angiotensin-binding protein in pig and rabbit liver during the course of studies aimed at identifying hepatic receptors for angiotensin II. After our publication of the cDNA sequence of the binding protein from pig (27), McKie *et al.* (18) pointed out the strong similarity between our sequence and that of rat EP 24.15 (TOP) which had been determined by Pierotti *et al.* (4, 19). We then obtained a second cDNA clone which was very similar to but clearly different from that of the cDNA for the binding protein, and we showed that the second clone represented the pig homolog of rat EP 24.15 (TOP) (20). The angiotensin-binding protein, although originally identified as a binding protein, did indeed have thiol- and metal-dependent oligopeptidase activity (20). At that time, therefore, the binding protein appeared to represent a new member of the thimet oligopeptidase family since the amino acid sequence of oligopeptidase M or EP 24.16 (MOP) from no mammalian species had yet been determined. 3) Kawabata *et al.* (28, 29) isolated an endopeptidase and the corresponding cDNA clone as a candidate for an enzyme responsible for the post-transcriptional processing of  $\gamma$ -carboxyglutamic acid-containing blood coagulation factors. They failed to notice the strong similarity to our binding protein, which was later pointed out by McKie *et al.* (18). 4) Checler *et al.* (20, 31) demonstrated the presence of a novel proteolytic activity capable of inactivating neuropeptides. They purified the peptidase from rat brain synaptic membranes and characterized it (32). The enzyme, termed neurolysin or endopeptidase 24.16, was shown to be distinct from EP 24.15 (TOP) and neprilysin (also known as enkephalinase or endopeptidase 3.4/24.11) and to have a relatively broad substrate-specificity and tissue distribution. Recent determination of its amino acid sequence by cDNA cloning clearly indicated that neurolysin is identical to the three enzymes mentioned above (33). Thus, four separate lines of research have converged in the discovery of a single new member of the thimet oligopeptidase family. In this report we use the abbreviation EP 24.16 (MOP) for this protein, whose identity has been only recently established.

EP 24.15 (TOP) and EP 24.16 (MOP) are very similar in terms of size and enzymatic properties: both are intracellular proteins of 78–80 kDa, consisting of about 680–700 amino acids, and their sequences are 65% homologous (20). They are, however, clearly distinguishable in several respects. For example, they have different specificities for inhibitors, different immunoreactivity, and different cleavage-site specificities. EP 24.15 (TOP) hydrolyzes neuropeptides exclusively at the Arg-Arg bond whereas EP 24.16 (MOP) cleaves it at the Pro-Tyr bond (16, 24, 32). Another difference is found in the subcellular localizations of these enzymes. EP 24.15 (TOP) is found in the cytosol while EP 24.16 (MOP) is found in both the cytosolic and mitochondrial compartments. How can the product of a single gene be localized to more than one intracellular compartment? To answer this question and to characterize evolutionary relationships among the members of the thimet oligopeptidase family, we investigated the structural organization of the pig genes for EP 24.15 (TOP) and EP 24.16 (MOP) and of their 5'-proximal flanking regions. We discovered six species of mRNA for EP 24.16 (MOP) that are generated from one single gene as a result of the utilization of alternative sites for the initiation of transcription. The six species of mRNA can be classified into two categories: those containing an additional sequence that encodes a mitochondrial targeting sequence and those that lack such a sequence. The use of different promoters for the eventual targeting of proteins to appropriate subcellular compartments appears to be a useful mechanism for adjusting

levels of local concentrations of proteins that function at different intracellular sites in response to the physiological requirements of the cell.

#### EXPERIMENTAL PROCEDURES

**Identification of 5'-Terminal Sequences of cDNAs for EP 24.16 (MOP)—**The 5'-ends of cDNAs for EP 24.16 (MOP) were cloned with the 5'-RACE (rapid amplification of cDNA ends) system (CLONTECH, Palo Alto, CA). Two  $\mu$ g of poly(A)<sup>+</sup> RNA, isolated from pig liver (27), were reverse-transcribed with a specific primer for the cDNA for pig EP 24.16 (MOP), 5RA-1 (5'-GTCTACCATGGTCCGTTCC-3'), and avian myeloblastosis virus reverse transcriptase. The first-strand cDNA was ligated at the 3'-end with an anchor (5'-CACGAATTCACTATCGATTCTG-GAACCTTCAGAGG-3') by T4 RNA ligase. A nested specific primer for the cDNA for EP 24.16 (MOP), 5RA-2 (5'-CCGCTCTACACCCCTACTTC-3'), was used with an anchor primer (5'-CTGGTTCGGCCCCACCTCT-GAAGGTCCAGAACATCGATAG-3') for amplification of the 5'-ends of the cDNAs by polymerase chain reaction. The products of polymerase chain reaction were fractionated on a 3% agarose gel, and fragments of 300–650 bp were isolated and cloned into pBluescript II (Stratagene, La Jolla, CA). Positive clones were identified by colony hybridization, with the  $^{32}$ P-labeled EcoRI-EcoRV 592-bp fragment of PAB-L1 (27) as probe, and sequenced.

**Sequencing of DNA—**DNA was sequenced by the dideoxy chain termination method of Sanger *et al.* (34) with double-stranded plasmids as templates. Termination reactions were performed with SequiTherm DNA polymerase (Epicentre Technologies, Madison, WI) and  $^{32}$ P-labeled M13 universal or reverse primer (LI-COR, Lincoln, NE). The products were analyzed with a DNA sequencer (model 4000; LI-COR). Sequences were organized and analyzed with GENETYX-MAC program (Software Development, Tokyo, Japan).

**Isolation of Genomic Clones for Pig EP 24.16 (MOP) and EP 24.15 (TOP)—**A pig liver genomic library constructed in  $\lambda$ EMBL3 SP6/T7 (CLONTECH) was screened with the 2.7-kilobase EcoRI-EcoRI fragment of a cDNA for EP 24.16 (MOP) clone (PAB-L1) (27), or with the 2.5-kilobase pair EcoRI-EcoRI fragment of a cDNA clone for EP 24.15 (TOP) (PABH-L7); (20), both of which had been labeled with [ $\alpha$ - $^{32}$ P]dCTP (Amersham, Little Chalfont, UK) with a random priming kit (Takara, Kyoto, Japan). Phage clones ( $2 \times 10^6$ ) were plated at a density of 30,000 plaque-forming units per  $135 \times 95$ -mm plate on *E. coli* NM538, from which duplicate replications were made on cellulose-nitrate filters (Schleicher & Schuell, Dassel, Germany) and allowed to hybridize with the  $^{32}$ P-labeled probe in a solution of 6  $\times$  SSPE (1  $\times$  SSPE is 0.15 M NaCl, 15 mM  $\text{Na}_2\text{PO}_4$ , pH 7.0, 1 mM EDTA), 50% formamide, 0.1% SDS, and 5 $\times$  Denhardt's solution at 42 °C for 16 h. The filters were rinsed twice at room temperature in 2  $\times$  SSC (1  $\times$  SSC is 0.15 M NaCl, 15 mM sodium citrate, pH 7.0) that contained 0.1% SDS and washed twice at 60 °C in 1  $\times$  SSC that contained 0.1% SDS for 1 h. Positive plaques were identified by autoradiography and purified by the additional rounds of screening.

**Restriction Mapping of  $\lambda$  Phages—**Positions of the EcoRI, SacI, and *Xba*I restriction sites in genomic clones were determined by complete or partial digestion with restriction enzymes and subsequent Southern blot analysis. UV irradiation and formation of pyrimidine dimers were used for preparation of incompletely digested genomic clones.  $\lambda$ EMBL3 SP6/T7 contains two unique *Sfi*I or *Sal*I sites and bacteriophage promoters (SP6 and T7) that flank the insert. Arms were separated with *Sfi*I or *Sal*I from the inserts, which still contained promoter sequences at the both ends. DNA samples were UV-irradiated for 0 or 20 min with UV Stratalinker 2400 (Stratagene) in 10 mM Tris, pH 7.5, 10 mM  $\text{MgCl}_2$ , and 1 mM dithiothreitol. UV-irradiated samples (600 ng) were digested incompletely with EcoRI, SacI, or *Xba*I (10 units) for 1 h at 37 °C, fractionated on a 0.7% agarose gel, and transferred to nylon membranes (Magnigraph; MSI, Westboro, MA). A set of filters was prepared and allowed to hybridize with end-labeled oligoprobes for T7 or SP6 promoter sequence, for 14 h at 37 °C in 6  $\times$  SSC, 5  $\times$  Denhardt's solution, 0.5% SDS, and 100  $\mu$ g/ml herring sperm DNA. The filters were washed twice in 1  $\times$  SSC, 0.1% SDS at 42 °C for 30 min, exposed to imaging plates, and analyzed with a Bioimage Analyzer (model BAS 2000; Fuji Film, Tokyo, Japan).

**S1 Nuclease Protection Assay—**Three primers, namely, 108L (CAAGCCCTGCGGCGGCCTAGCAAAGGAGGCAACAG) for exon 1; 107L (GGTGTCCCTCGGGTAGACCATGTGGCTGTAGAA) for exon 2; and 106b (GTCTCTCCATGAGAATGCTCCT) for exon 3, were designed for the synthesis of single-stranded antisense DNA probes that would protect pig 5'-ends of mRNAs for EP 24.16 (MOP). Ten pmol of each primer were labeled with [ $\gamma$ - $^{32}$ P]ATP (Amersham) by polynucle-

**Fig. 1. Nucleotide sequences of the six types of 5'-end of cDNAs for pig EP 24.16 (MOP).** The six different sequences revealed by 5'-RACE are aligned. Candidates for codons for initiation of translation are boxed and labeled *M1*, *M2*, and *M3*, respectively. These initiation sites are aligned with the consensus sequence for sites of initiation of translation (GCCA/GCCATGG) (45) which is shown in *italics*. Three nucleotide replacements, due possibly to allelic polymorphism, were found in exon 2 (*underlined*). Exon boundaries, determined by comparison of genomic and cDNA sequences (Fig. 4D), are indicated by *vertical arrows*.

otide kinase (Takara) and used for the synthesis of probes. End-labeled primers were annealed with 5  $\mu$ g of plasmid DNA that contained genomic fragments of the pig gene for EP 24.16 (MOP) (*Apal-Xba*I 837-bp fragment of APAB-G33 for exons 1 and 2; *Bgl*II-*Eco*RI 923-bp fragment of APAB-G32 for exon 3); and antisense probes were synthesized with T7 DNA polymerase (Pharmacia, Uppsala, Sweden). The 3'-ends of the probes were digested with restriction enzymes (*Sma*I for exon 1, *Bss*HII for exon 2, and *Bgl*II for exon 3), fractionated by electrophoresis on a 5% polyacrylamide gel that contained 7 M urea and exposed to x-ray film. Probes were detected as bands of the expected mobility and extracted in 0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% SDS, and 10  $\mu$ g/ml yeast tRNA at 37 °C for 12 h. Extracted probes were precipitated in ethanol, and probes (1  $\times$  10<sup>5</sup> cpm each) were annealed with 5  $\mu$ g of poly(A)<sup>+</sup> RNA from pig liver or with 10  $\mu$ g of yeast tRNA, as a control, for 12 h at 30 °C in 80% formamide, 40 mM PIPES, pH 6.4, 1 mM EDTA, and 400 mM NaCl. Non-annealed nucleic acids were digested with S1 nuclease (Boehringer Mannheim, Mannheim, Germany) at a final concentration of 1,000 units/ml in 0.28 M NaCl, 0.05 M sodium acetate, pH 4.5, 4.5 mM ZnSO<sub>4</sub>, and 20  $\mu$ g/ml denatured herring sperm DNA. The protected fragments were purified by extraction with phenol/chloroform and precipitation with ethanol, and electrophoresed in 5% polyacrylamide gels containing 7 M urea. Gels were dried and exposed to imaging plates for 48 h. Images were analyzed with the Bioimage Analyzer.

Production of Antiserum against Purified Pig EP 24.16 (MOP)—Rabbits were injected subcutaneously with 75 µg of purified pig EP 24.16 (MOP) (formerly referred to as soluble angiotensin-binding protein, sABP (27)) in complete Freund's adjuvant. Booster injections with 75 µg of purified protein in incomplete Freund's adjuvant were given 2, 4, and 6 weeks after the initial injection. Rabbits were bled 10 days after the fourth injection.

**Construction and Expression of cDNAs for Isoforms of Pig EP 24.16 (MOP)—**Six plasmids, pcDNA3-MOP1 (exon 1-[5-16]); -MOP1' (exon 1-4-[5-16]); -MOP2 (exon 2-[5-16]); -MOP2' (exon 2-4-[5-16]); -MOP3 (exon 3-[5-16]); and -MOP3' (exon 3-4-[5-16]), were constructed for expression analysis. For pcDNA-MOP1, a 2732-bp EcoRI-EcoRI fragment of PAB-L1 (27), which contained the entire open reading frame of type 1 cDNA for EP 24.16 (MOP), was subcloned to pcDNA3 (Invitrogen, San Diego, CA). For the other plasmids, PAB-R5, -R302, -R305, -R8, and -R1 (Fig. 1), which encoded only the 5'-ends of type 1', 2, 2', 3, and 3' cDNAs, respectively, were digested with *Alu*NI at their 3'

termini, ligated with the 2291-bp *Alu*N1-EcoRI fragment of PAB-L1 and subcloned into pcDNAS. COS-1 cells were maintained in Dulbecco's modified Eagle's medium (Life Technologies, Inc., Gaithersburg, MD) that contained 10 mM HEPES, pH 7.2, 10% fetal bovine serum, 50 units/ml penicillin, and 50  $\mu$ g/ml streptomycin, in a controlled atmosphere of 5% CO<sub>2</sub> in air at 37 °C. Approximately  $6 \times 10^6$  cells were electroporated with 20  $\mu$ g of each plasmid at 220 V at a capacitance setting of 960 microfarads in a Gene Pulser apparatus (Bio-Rad) and harvested 48 h after electroporation.

**Subcellular Fractionation of Cells and Western Blotting**—All steps were performed at 4 °C. Cells were washed by centrifugation in Dulbecco's phosphate-buffered saline (2.7 mM KCl, 138 mM NaCl, 1.2 mM KH<sub>2</sub>PO<sub>4</sub>, and 8.1 mM Na<sub>2</sub>HPO<sub>4</sub>, pH 7.4) at 700 rpm for 2 min. Approximately 2 × 10<sup>7</sup> cells were suspended in 1 ml of 2.5 M sucrose and homogenized for 2 min. Nuclear fractions were removed by centrifugation at 3,000 rpm (700 × g) for 10 min, and supernatants were centrifuged at 9,200 rpm (7,000 × g) for 10 min to recover mitochondrial fractions as pellets. Mitochondrial fractions were washed twice by centrifugation at 25,000 rpm (24,000 × g) for 10 min. The post-mitochondrial supernatants were centrifuged at 50,000 rpm (105,000 × g) for 100 min, and the pellets (microsomes) and supernatants (cytosol) were recovered. The concentration of protein in each fraction was determined with the BCA protein assay reagent (Pierce). Five µg of each protein sample were fractionated by SDS-PAGE (10% polyacrylamide) in standard glycine running buffer (192 mM glycine, 25 mM Tris, and 0.1% SDS) or high-resolution running buffer (492 mM glycine, 75 mM Tris, and 0.1% SDS). The separated proteins were transferred to a polyvinylidene difluoride membrane (ATTO, Tokyo, Japan) and probed with 2,000-fold diluted rabbit antiserum against pig EP 24.16 (MOP). Bound antibodies were detected with alkaline phosphatase-conjugated second antibodies, with 4-nitro blue tetrazolium chloride and 5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP) as chromogen.

## RESULTS

*Differences among 5'-Terminal Sequences of mRNAs for EP 24.16 (MOP)*—To delineate the complete structure of the gene for EP 24.16 (MOP), we determined the 5'-end of the corresponding mRNA by rapid amplification of 5'-ends of cDNA (5'-RACE) using preparations of poly(A)<sup>+</sup> RNA from pig liver. More than five prominent bands of fragments of 340–630 nu-

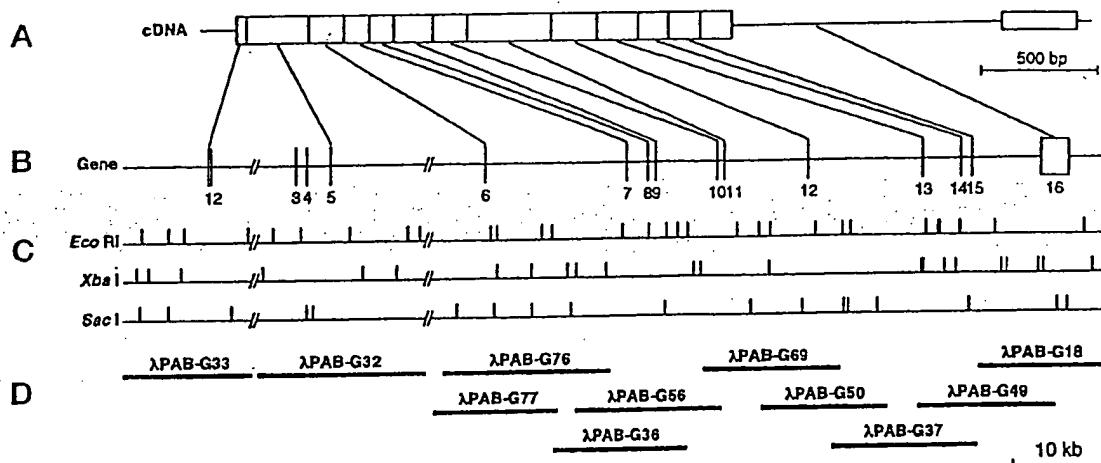


FIG. 2. Structure of the cDNA for pig EP 24.16 (MOP) in relation to the structure of the gene (A), exon-intron organization of the gene (B), restriction maps for *Eco*RI, *Sac*I, and *Xba*I (C), and the relative locations of genomic clones used for the analysis (D). The structure of cDNA for pig EP 24.16 (MOP) is described elsewhere (27).

cleotides were obtained. DNA sequencing of these fragments revealed the presence of several mRNAs whose sequences were completely different from others starting 28 nt upstream from the ATG initiation codon (Fig. 1; M3). These results suggested that usage of alternative promoters and exons might be involved in the generation of the observed mRNA diversity. To determine the precise molecular mechanism responsible for generation of such heterologous mRNAs, we isolated and characterized the pig gene for EP 24.16 (MOP), which had previously been shown to be present as only a single copy (27).

**Structural Organization and 5'-Untranslated Exons of EP 24.16 (MOP) Gene**—To isolate the pig gene for EP 24.16 (MOP), we screened approximately  $2 \times 10^6$  independent plaques of a pig genomic library in λEMBL3 (CLONTECH) using the PAB-L1 cDNA clone (27) as the probe. We isolated and mapped more than 50 clones, and then we subcloned and sequenced the phage fragments for the identification of exons. The exon-intron organization of the pig gene for EP 24.16 (MOP) was deduced from an analysis of 11 independent clones, each of which contained part of the gene (Fig. 2). The gene extends over 100 kilobases and contains 16 exons and 15 introns of various sizes (Figs. 2B and 4B). All the introns have typical splice donor and acceptor boundaries (Fig. 4D) (35).

Comparison of the nucleotide sequences of the genomic clones λPAB-G32 and λPAB-G33 (Fig. 2D) with those of the products of 5'-RACE (Fig. 1) allowed us to identify the alternatively spliced leader exons (Fig. 2, A and B, and 6A). Six distinct species of mRNA for EP 24.16 (MOP) appeared to be generated by differential use of three sites for initiation of transcription located upstream of exons 1, 2, and 3, respectively, and by the alternative splicing of exon 4; exons 1, 2, and 3 are mutually exclusive (Fig. 6A). Exon 1 encodes a putative mitochondrial targeting sequence, (M)IVRCLSAARRLHR (Fig. 6D), which is rich in basic amino acids and can be expected to form an amphipathic helix (36). The common exons 5 through 16 are used to assemble the functional domain of the enzyme. The zinc-binding motif HEFGH is encoded by exon 12 (Fig. 4B). The extreme 3' exon, exon 16, encodes the last 44 amino acid residues, the termination codon, and the 3'-untranslated sequences that include three polyadenylation signals, a short interspersed repetitive element (SINE or PRE-1), and an AT repeat, all of which were identified previously by cDNA cloning (27).

There appears to be a "pseudo-exon" that encodes a protein that resembles a ribosomal protein (11.5 kDa, L44 (37)) in

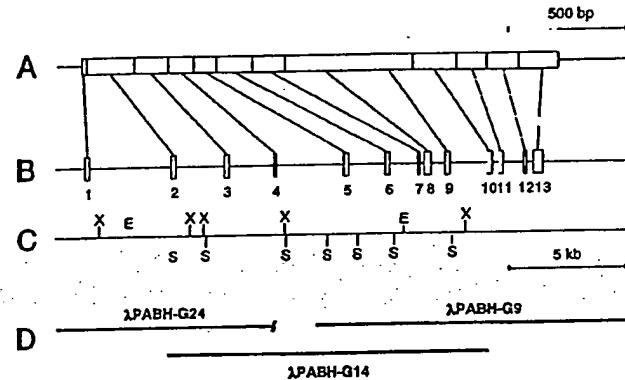


FIG. 3. Structure of the pig gene for EP 24.15 (TOP). A, structural relationship to the cDNA. The structure of cDNA for pig EP 24.15 (TOP) is described elsewhere (20). B, exon-intron organization. C, restriction maps. The following abbreviations are used for the restriction enzymes: E, *Eco*RI; S, *Sac*I; X, *Xba*I. D, relative locations of genomic clones.

reverse orientation (3' to 5') within the untranslated region of the 3'-most exon (Fig. 4B). The sequence encoding the homolog of ribosomal L44 is flanked by the direct repeat TGTTTTA-GAGAATT and has a poly(A) tract, suggesting that the pseudogene might have arisen as a result of retroposition.

**Structural Comparison of the Genes for EP 24.15 (TOP) and EP 24.16 (MOP)**—We wondered whether the complexity of organization of the gene for EP 24.16 (MOP) might be reflected in the genes for other members of the thimet oligopeptidase family and, to this end, we also characterized the gene for EP 24.15 (TOP). The gene for EP 24.15 (TOP) was isolated from the same pig genomic DNA library as that used for isolation of genes for EP 24.16 (MOP), and it was found to have a much simpler structure in its 5'-region (Figs. 3 and 4B). The gene exists as a single copy, as revealed by Southern blot analysis (data not shown); it spans approximately 45 kilobase pairs (Fig. 4A); and it is organized into 13 exons. The overall organization of the two genes is very similar with the exception of the length of introns and the 5'-leader and untranslated exons (Fig. 4B). For example, exons 2-12 of the gene for EP 24.15 (TOP) correspond precisely to exons 5-15 of the gene for EP 24.16 (MOP) and there is strong conservation of the respective exon-intron boundaries (Fig. 4D), suggesting evolution from a common ancestor. The zinc-binding motif HEFGH is encoded within

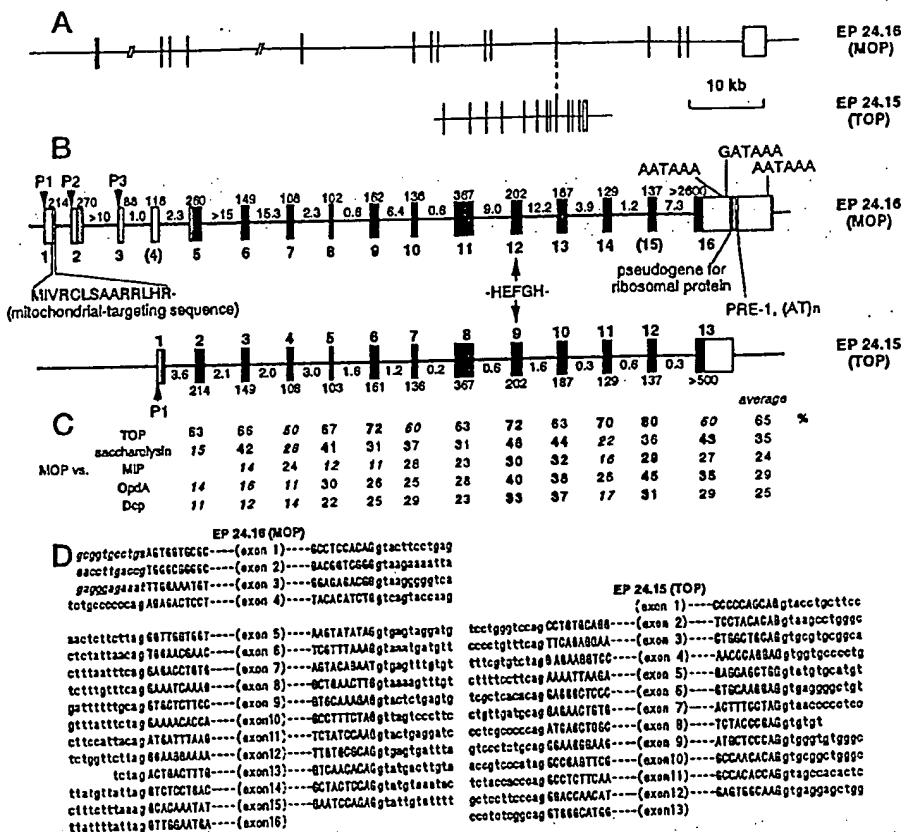


FIG. 4. Comparison of the structures of genes for EP 24.16 (MOP) and EP 24.15 (TOP). *A*, comparison of the lengths of pig genes for EP 24.16 (MOP) and EP 24.15 (TOP). The exons corresponding to the zinc-binding motifs (exon 12 of EP 24.16 (MOP) and exon 9 of EP 24.15 (TOP)) are aligned (dotted vertical line). *B*, comparison of the exon-intron organization of the gene for EP 24.16 (MOP) with that of EP 24.15 (TOP). Exons are indicated by *boxes* and are numbered in **bold** type. Coding regions for the peptidases are shown by *filled boxes* and amino-terminal extension sequences generated from the alternative sites of initiation of translation of EP 24.16 (MOP) are *hatched*. Numbers indicate the sizes of exons and introns in base pairs and kilobase pairs, respectively. *C*, the extent of sequence homology between the product of the gene for EP 24.16 (MOP) and related peptidases. Coding exons of the gene for EP 24.16 (MOP) (exon 5–16) were converted into deduced amino acid sequences and compared with those of other peptidases, such as pig EP 24.15 (TOP), saccharolysin from *S. cerevisiae* (10), rat mitochondrial intermediate peptidase (MIP) (8), oligopeptidase A (OpdA) from *E. coli* (52), and dipeptidyl carboxypeptidase (Dcp) from *E. coli* (53). The extent of homology is indicated, as a percentage, below the corresponding exons. **Bold numbers** indicate relatively strongly conserved regions, with homology scores more than five points above the average. *Italic numbers* indicate poorly conserved regions, with homology scores more than five points lower than the average. *D*, nucleotide sequences of exon-intron junctions. The exons are shown in *capital letters* and the introns in *lowercase letters*. The 5'-flanking regions of three sites of initiation of transcription are shown in *italics*.

exon 9. The 3'-terminal exon 13 is composed of a short coding sequence, the termination codon, and the entire 3'-untranslated sequence. The promoter region of the gene for EP 24.15 (TOP) lacks the TATA box but contains several putative binding sites for ubiquitous factors including one CCAAT box, three Sp1 sites, one NF-1 site, one AP-1 site, and two AP-2 sites (data not shown).

*Identification of Three Major Sites of Transcription Initiation*—Characterization of the 5'-ends of mRNAs for EP 24.16 (MOP) by 5'-RACE (Fig. 1) revealed the presence of multiple sites for initiation of transcription, at least one each in the upstream of exons 1, 2, and 3. To determine the transcription start sites, we performed S1 nuclease mapping using poly(A)<sup>+</sup> RNA from pig liver and using three probes, which were complementary to exons 1, 2, and 3, respectively. The locations of these probes are indicated in Fig. 5A. As we had expected, we found three sites (Fig. 5A): one located 172 nt upstream of the first Met (ATG) codon of exon 1; another located at 106 nt upstream of the Met codon of exon 2; and the third located 23 nt downstream of the TATA box close to exon 3.

*Putative Sites for Binding of Transcription Factor Near Sites of Initiation of Transcription*—Inspection of the sequence of 5'-flanking regions of exons 1, 2, and 3, which we designated

promoter regions P1, P2, and P3, respectively (Fig. 4B), revealed potential *cis*-acting DNA elements (Fig. 5, B and C). Promoter regions 1 and 2 are very GC-rich and lack the TATA and CAAT boxes that are typical of eukaryotic class II promoters; promoter 3 contains a conserved TATA box, which begins 29 nt upstream of the previously identified 5'-end of exon 3. The sequences upstream of exons 1 and 2 contain several putative binding sites for transcription factors AP-2 and Sp1. AP-2 mediates enhanced transcription as a result of stimulation by the protein kinase C, cAMP-dependent protein kinase A, and retinoic acid (38–40); Sp1 is a protein that binds to the GC box specifically and is often involved in the regulation of so-called housekeeping genes (41). The upstream region of exon 3 includes consensus binding sites for the transcription factors Myb (product of the myeloblastosis oncogene), AP-1, and GATA-1. The presence of multiple binding sites for hematopoiesis-specific factors is intriguing: Myb has been demonstrated to be important in the control of the proliferation and differentiation of hematopoietic cells (42), while GATA-1 was originally found as an erythroid-specific factor (43). The Myb-binding site immediately downstream of the TATA box for exon 3 is of particular interest since such a juxtaposed arrangement of a TATA box and a Myb target sequence was recently demon-

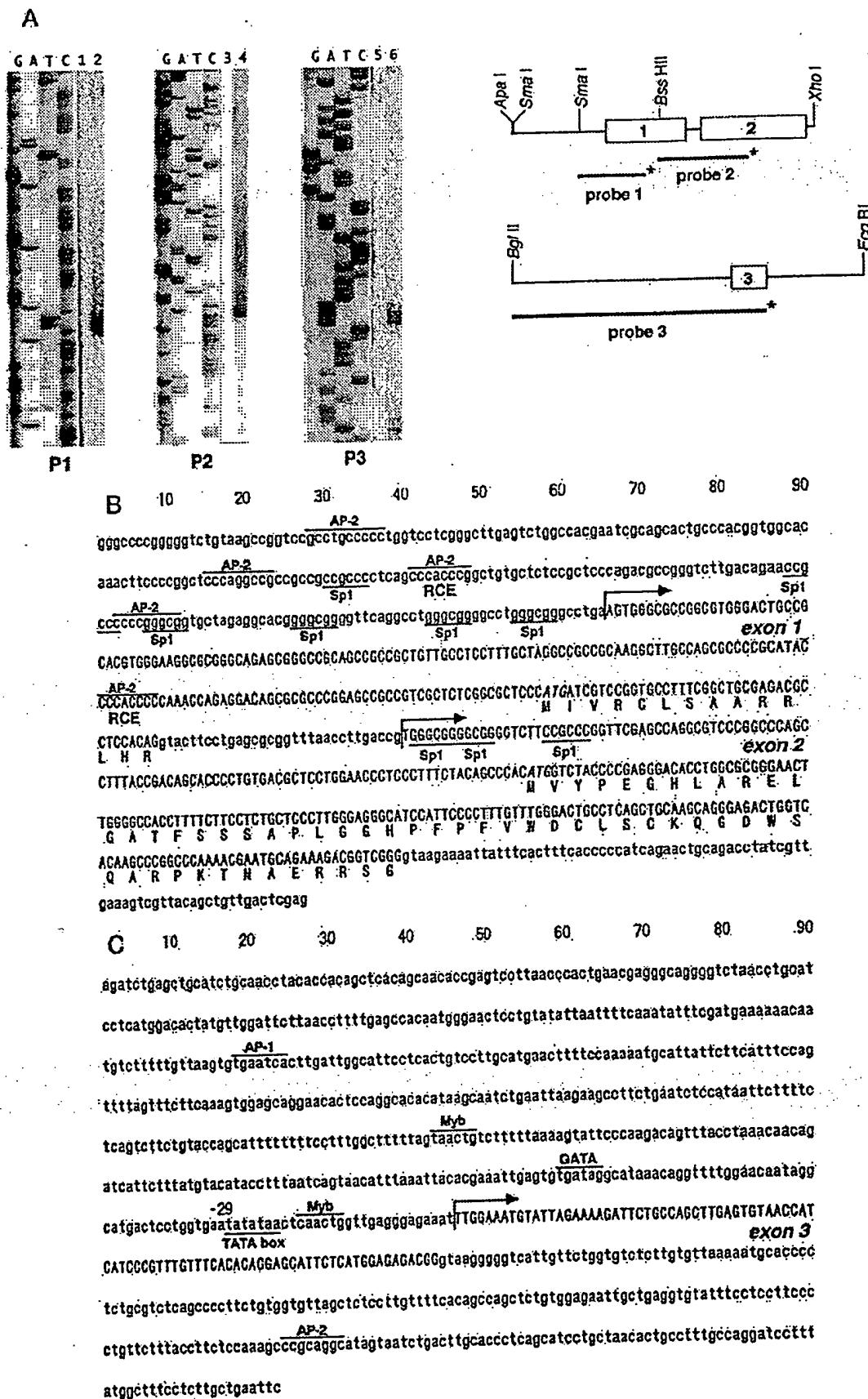


FIG. 5. Identification of site of initiation of transcription of the pig gene for EP 24.16 (MOP) by S1 nuclease mapping (A) and sequences of promoter regions (B and C). A, 5  $\mu$ g of poly(A)<sup>+</sup> RNA from pig liver (lanes 2, 4, and 6) or yeast tRNA (lanes 1, 3, and 5) were

strated to serve as a Myb-suppressible promoter (44).

**Differential Subcellular Localization of EP 24.16 (MOP) Directed by Alternatively Generated Species of mRNA**—The results described above suggest that the organization of the 5'-region of the genes for EP 24.16 (MOP) is unusually complex and that six mRNA species with different 5' termini are generated as a consequence of the use of separate promoters (Fig. 5) and the splicing of the 5'-leader exons 1 through 3 (in a mutually exclusive manner) and of exon 4. The cDNA sequences corresponding to the six species of mRNA are shown schematically in Fig. 6A, and they were used for the expression experiments described below. It should be noted that exon 1 has an in-frame ATG codon (designated M1), when connected directly to exon 5, and the open reading frame in exon 1 encodes a putative signal peptide for import into mitochondria; exon 2 also has an in-frame ATG codon (M2) in an appropriate context for the initiation of translation (Fig. 1) (45) and the open reading frame predicts an enzyme with 64 more amino acids at its amino terminus than the product generated by the open reading frame that starts with an ATG codon (M3) in the common exon 5 (Figs. 1 and 6A). The fact that exon 1 could encode an amino-terminal leader sequence for targeting to mitochondria strongly suggests that, upon selection or elimination of the sequence of exon 1 via differential utilization of the multiple promoters, the subcellular localization of the products of the gene for EP 24.16 (MOP) is strictly and efficiently controlled. To confirm this possibility, we carried out the following experiments.

The six cDNA constructs depicted on the right side of Fig. 6A (labeled types 1 through 3 and 1' through 3') were inserted separately into the mammalian expression vector pcDNA3 and used to transfect COS-1 cells. Then subcellular organelles were isolated from the transfectants and the levels of EP 24.16 (MOP) in these organelles were examined by Western blotting (Fig. 6, B and C). The type 1 (1-[5-16]) construct directed the synthesis of EP 24.16 (MOP) that was targeted to mitochondria (Fig. 6, B, lane 3, and C, MOP(M1b)); the mitochondrial enzyme was slightly smaller than the unprocessed precursor that remained, as a consequence of overexpression of the protein, in the cytosol (Fig. 6, B, lane 2, and C, MOP(M1a)). This difference in size indicates that the amino-terminal mitochondrial targeting sequence is cleaved after translocation of the protein into mitochondria. Type 1' (1-4-[5-16]), in which the connection between exons 1 and 5 is interrupted by insertion of exon 4 which includes a stop codon (Fig. 6D), yielded only the cytosolic form of EP 24.16 (MOP) generated from the ATG initiation codon (M3) in exon 5 (Fig. 6B, lanes 5-8). Type 2 (2-[5-16]) allowed the synthesis of an amino-terminally extended cytosolic form (Fig. 6B, lane 10, upper band). Again, as seen with type 2' (2-4-[5-16]), insertion of exon 4 generated a stop codon and only the short cytosolic form was expressed (Fig. 6B, lanes 9-16). With type 1 and type 2, products of translation from the ATG codon in exon 5 (M3) were also detected (Fig. 6, B, lane 10, lower band, and C, MOP(M3)), suggesting that these mRNAs generate two isoforms of the protein by alternative usage of codons for the initiation of translation (M1 and M3 for type 1 and M2 and M3 for type 2). The constructs having exon 3 as the 5'-leader exon (types 3 and 3') produced only the cytosolic form

of the enzyme (Fig. 6B, lanes 17-24), as expected from the fact that exon 3 contains no in-frame ATG codon.

## DISCUSSION

In this study, we demonstrated the heterogeneity at the 5'-end of the mRNA for EP 24.16 (MOP). Moreover, we showed that the heterogeneity is generated by alternative usage of promoters and splicing of multiple 5'-leader and untranslated exons and that it is responsible for the differential subcellular localization of the products of translation.

**Targeting of Proteins to Different Subcellular Locations by Alternative Usage of Promoters: EP 24.16 (MOP) Represents the First Example of Such a Mechanism for Intracellular Peptidases/Proteinases**—Proteins, after their synthesis, must be delivered to their sites of action. Delivery is usually accomplished with the help of terminal or internal targeting sequences. Sequences for the targeting proteins to the following sites have been identified: mitochondria, endoplasmic reticulum, lysosomes, nuclei, and peroxisomes.

The presence of a putative mitochondrial targeting sequence at the amino terminus of the precursor to EP 24.16 (MOP) was first deduced by Serizawa *et al.* (24) from the potential ability of this sequence to form an amphipathic  $\alpha$ -helix with a hydrophobic and a positively charged face of the type expected for a mitochondrial leader sequence (36, 46). This scenario explains the presence of EP 24.16 (MOP) in mitochondria. The enzyme is, however, known also to be present in the cytosol and, prior to the present study, the mechanism responsible for this distribution of EP 24.16 (MOP) has remained unclear. Discovery of 5'-end variants of the mRNA for EP 24.16 (MOP) by the 5'-RACE technique led us to investigate the genetic basis for such diversity. Through an analysis of the structure of the gene, which led to the identification of the three 5'-leader exons that are selected, in a mutually exclusive manner, by use of alternative promoters and splicing, we provided the following resolution of this problem (Fig. 7). If promoter 1 is used, the mitochondrial isoform of EP 24.16 (MOP) is generated by splicing of exon 1, which has a sequence that encodes a signal for transport to mitochondria, to exon 5, which is the beginning of the common translated region that encodes the mature portion of the protein (type 1 in Fig. 6, A and D). The precursor form (704 amino acid residues) with the mitochondrial targeting sequence is processed to the mature mitochondrial form of 667 residues (Fig. 7). The type 1 transcript can also yield the cytosolic form of 681 amino acids when the M3 site of initiation of translation is used instead of the M1 site. If promoter 3 is used and exon 3, which lacks an in-frame ATG codon, is joined to exon 5 (type 5 in Fig. 6, A and D), the cytosolic isoform is produced from the ATG initiation codon in exon 5. The use of promoter 2, which directs the synthesis of a cytosolic variant, is discussed below.

Similar scenarios have been reported for several other enzymes that are known to occur and function in more than one subcellular compartment (for a recent review, see Ref. 47). Typical examples are the histidine and valine tRNA synthetases of *Saccharomyces cerevisiae* that are involved in protein synthesis in the cytosol and the mitochondria (48, 49). In these cases, two types of transcript (long and short) are produced by

used. Sequencing with the same oligonucleotides was used for calibration of mobilities. Strategies for the preparation of single-stranded antisense DNA probes are shown on the right. B, the sequence of the 837-bp *Apa*I-*Xba*I fragment of pig genomic DNA that contained exons 1 and 2. Sites of initiation of transcription are indicated by arrows. Exon 1 and exon 2 are located in a very small region with a GC content of 67%. Nine binding sequences for Sp1, five binding sequences for AP-2, and two Rb control elements (RCE) are indicated. Capital letters represent exons and the deduced amino acid sequences are shown. C, the sequence of the 923-bp *Bgl*II-*Eco*RI fragment of pig genomic DNA that contained exon 3. A site of initiation of transcription is indicated by an arrow. The sequence includes a TATA box at position -23 relative to the site of initiation of transcription of exon 3, three binding sites for Myb, one for AP-1, and one for AP-2. The binding site for Myb site near the TATA box is very similar to the sequence that is found in the *c-erbB-2* promoter that has been shown to suppress this gene (44).

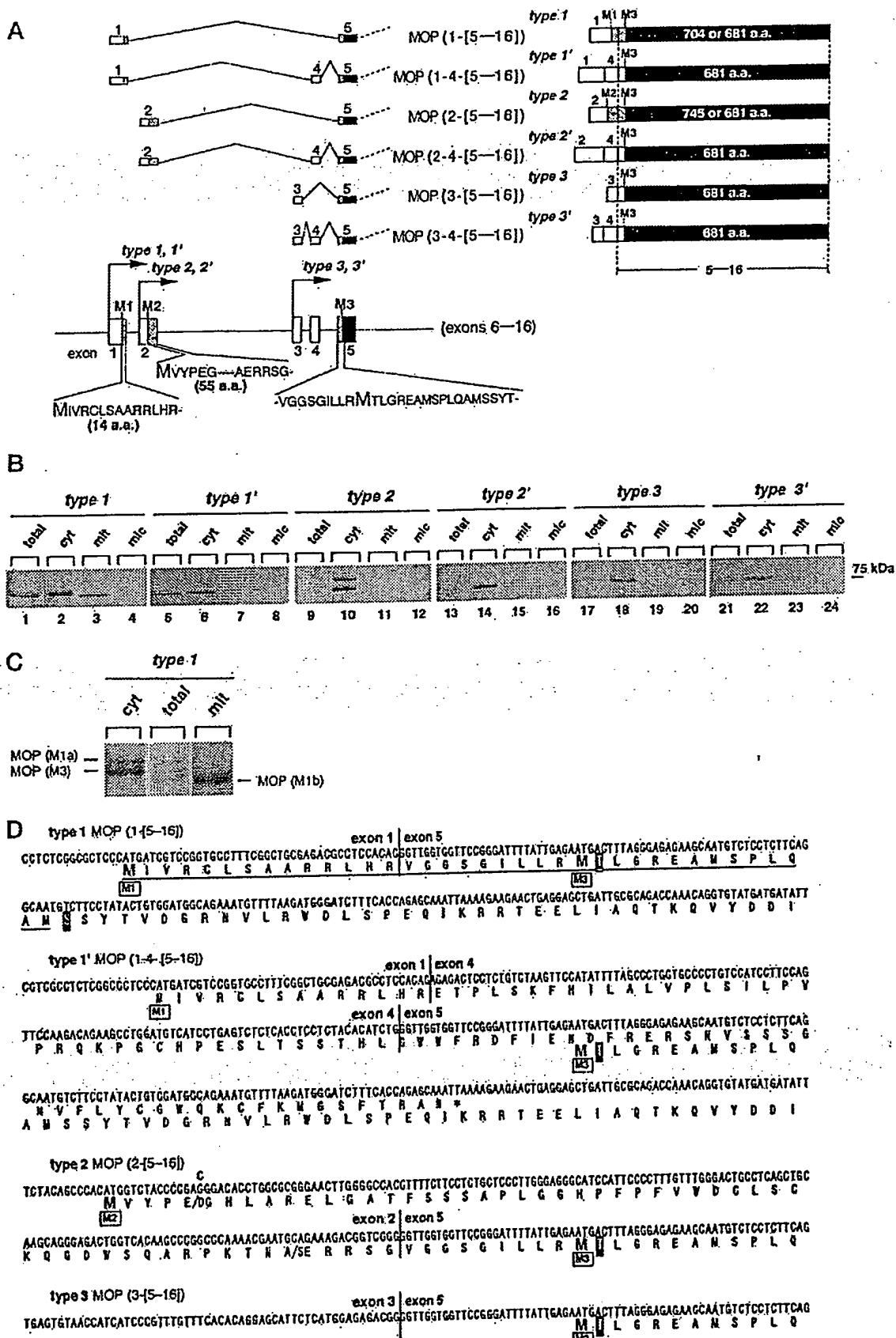


FIG. 6. Generation of cytosolic and mitochondrial forms of EP 24.16 (MOP) from a single gene by alternative usage of three promoters (P1-P3) and three codons for initiation of translation (M1-M3). A, schematic representation of the six isoforms of the mRNA for EP 24.16 (MOP) and the organization of the 5'-region of the gene for EP 24.16 (MOP), showing how the various isoforms are generated. Exons

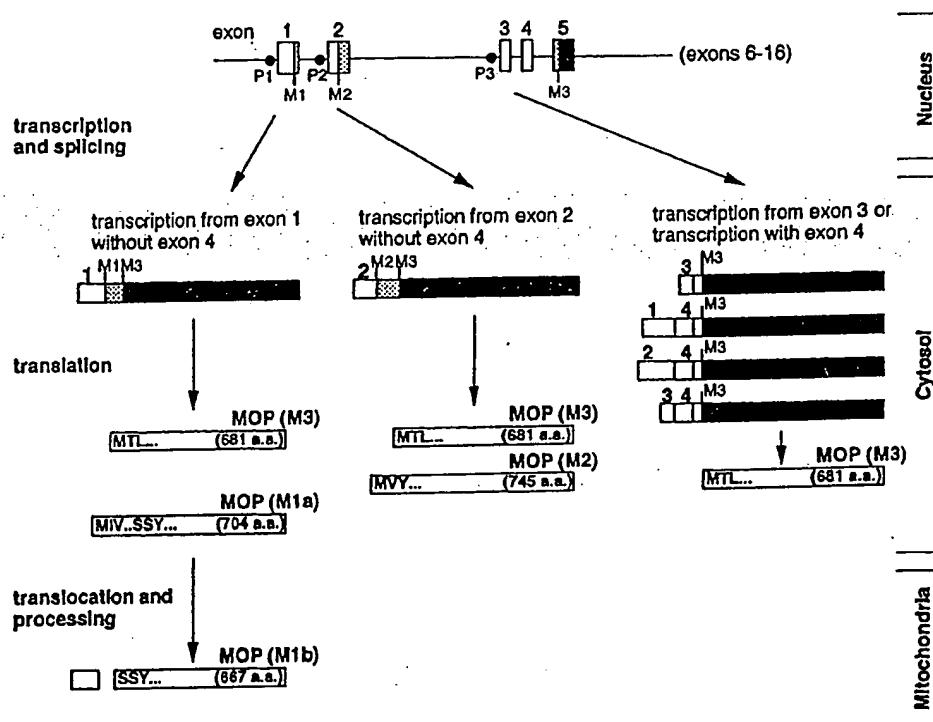


FIG. 7. Schematic representation of the mechanism for generation of multiple forms of EP 24.16 (MOP). Three promoters (P1, P2, and P3) and three codons for the initiation of translation (M1, M2, and M3) are used for the expression of EP 24.16 (MOP). Activation of both P1 and M1 is necessary for expression of EP 24.16 (MOP) in mitochondria. Insertion of exon 4 prevents the translocation of EP 24.16 (MOP) to mitochondria.

alternative usage of promoters, and the long transcript yields the mitochondrial isoform exclusively, while the short transcript yields the cytosolic enzyme. In this way, adjustment of the levels of the proteins to the needs of each compartment is possible. Although the biological significance of this mechanism in the present case is not immediately apparent since the true substrates of the enzyme have not yet been identified, the general regulation of expression of the gene for an oligopeptidase by transcription regulatory factors and the unique regulation, as reported herein, of targeting of the product by use of alternative promoters seems to provide a powerful method by which cells can modulate the concentration of specific peptides in certain intracellular compartments to reflect the metabolic state of the cell.

**Complex Organization of the Pig Gene for EP 24.16 (MOP)**—As compared with the gene for EP 24.15 (TOP), another member of the thimet oligopeptidase family, the gene for EP 24.16 (MOP) have quite a complicated structure in its 5'- and 3'-regions. The two genes do, however, exhibit extensive

similarity in the regions that encode the mature proteins, which consist of 11 exons, namely, exons 2 through 12 in the case of the gene for EP 24.15 (TOP) and exons 5 through 15 in the case of the gene for EP 24.16 (MOP). The similarity suggests that these two genes and, probably, the genes for other members of this family were generated from an ancestral gene as distinct sequences as a consequence of gene duplication. The presence of a SINE in the 3'-untranslated region of the gene for EP 24.16 (MOP) suggests that insertion of a SINE after the gene duplication might have destabilized the gene for EP 24.16 (MOP) and stimulated extensive diversification of 5'- and 3'-regions by recruiting the entire gene for ribosomal protein L44 (in the reverse orientation) into the 3'-most exon (exon 16) and the 5'-leader exons into the 5'-flanking region by, perhaps, retroposition and gene conversion.

**Long and Short Forms of EP 24.16 (MOP)**—The use of promoter 2 of the gene for EP 24.16 (MOP) yields the type 2 transcript, which is predicted to have an amino-terminally elongated product (Fig. 7). Consistent with this prediction, we

are indicated by boxes and numbered (box patterns: black, coding regions for the peptidase; hatched, reading frames encoding the amino-terminal extensions; and white, non-coding regions). Three sites for initiation of transcription (Fig. 5) are indicated by arrows. The deduced amino acid sequences corresponding to the reading frames of exons 1, 2, and 5 are shown in single letter code, that contains translational initiation sites indicated by bold and large letters. Basic amino acid residues which are necessary for mitochondrial targeting sequences are indicated by bold letters. Alternative initiation of the transcription of exons 1-3 (P1-P3) and alternative splicing of exon 4 generate six isoforms of the mRNA. *B*, letters. Subcellular localization of the products of translation of isoforms of the cDNA for EP 24.16 (MOP) expressed in COS-1 cells. Six cDNA species, identified by 5'-RACE (Figs. 1 and panel A, of this figure), were expressed in COS-1 cells and the products were detected by Western blotting. Subcellular fractions were obtained by differential centrifugation, as follows: *cyt*, cytosol (100,000 × g supernatant); *mit*, mitochondria (7,000 × g pellet); *mic*, microsomes (100,000 × g pellet). Construction of cDNA used for expression is indicated by boxes on the right of panel (A). *C*, resolution of the mitochondrial form generated by use of the first site (M1) for initiation of translation; *M1b*, the processed mature form imported into mitochondria; *M3*, the cytosolic form generated by translation from the M3 site of initiation of translation (for details see Fig. 7). *D*, nucleotide and amino acid sequences of the type 1, 1', 2, and 3 isoforms. Three codons for initiation of translation, M1, M2, and M3, are present in exon 1, exon 2, and exon 6, respectively. The mitochondrial targeting sequence of EP 24.16 (MOP), containing six arginine residues (24), is underlined. Serizawa *et al.* (24, 54) determined the amino-terminal amino acid sequences of the isoforms of EP 24.16 (MOP) purified from mitochondria and the cytosol and showed that the mitochondrial form has a Ser residue at its amino terminus and that the major cytosolic form begins with Thr; the Ser and Thr residues are indicated by white lettering on a black background.

detected a long form of the protein in the cytosol of COS-1 cells transfected with the type 2 construct after SDS-PAGE and Western blotting (Fig. 6B, lane 10). Although the relative abundance of the corresponding mRNA, as estimated from the data after 5'-RACE, in the pig liver is low (<10%), the physiological significance of this form clearly merits further study. The roles of the extended amino-terminal region of 64 amino acid residues (Fig. 6D) could include stabilization of the enzyme, modulation of the substrate specificity, and/or mediation of interactions with other cytosolic proteins.

In our analysis, we also noticed the presence of a splice variant that lacked the sequence of exon 15.<sup>2</sup> This variant should encode a protein with a short and slightly different carboxyl-terminal tail. The functional significance of this variant and the tissue- and development-specific regulation of the splicing will be the subject of further research. The presence of at least two forms of EP 24.16 (MOP) has also been demonstrated in purified preparations of the enzyme from rabbit and pig liver (50, 51).

The type 1, type 2, and type 3 species of mRNA all have splice variants, designated type 1', type 2', and type 3', respectively, with an extra exon sequence (5'-untranslated exon 4, Fig. 7), but none of them results in a long open reading frame from the first initiation codon. It is unknown therefore, whether these variants have any biological significance. However, the possibility exists that the variant species of mRNA might contribute to the regulation of rates of translation. Alternatively, they might produce short peptides with as yet unidentified functions. In the case of the type 1' transcript, the insertion of the sequence of exon 4 provides a mechanism by which the synthesis of the mitochondrial isoform is suppressed even under conditions under which promoter 1 is active but there is no mitochondrial requirement for the oligopeptidase.

**Conclusion**—Analysis of the gene for EP 24.16 (MOP) revealed the very complex organization of the gene and the presence of a variety of transcripts generated by differential use of multiple sites of initiation of transcription and by alternative splicing of exons 2, 3, 4, and 15. In contrast to these complexities, a simple and definitive answer was obtained to the question of how the product of a single gene for EP 24.16 (MOP) is delivered to two different cellular compartments, namely, the cytosol and the mitochondria.

**Acknowledgments**—We thank Makoto Itakura, Takeshi Ihara, and Hiromi Hagiwara for helpful discussions, and Setuko Satoh and Kazuko Tanaka for expert secretarial and technical assistance.

#### REFERENCES

- Bond, J. S., and Beynon, R. J. (1985) *Int. J. Biochem.* **17**, 565-574
- Tislijar, U. (1993) *Biol. Chem. Hoppe Seyler* **374**, 91-100
- Barrett, A. J., Brown, M. A., Dando, P. M., Knight, C. G., McKie, N., Rawlings, N. D., and Serizawa, A. (1995) *Methods Enzymol.* **248**, 529-556
- Pierotti, A., Dong, K. W., Glucksman, M. J., Orlowski, M., and Roberts, J. L. (1990) *Biochemistry* **29**, 10323-10329
- Orlowski, M., Reznik, S., Ayala, J., and Pierotti, A. R. (1989) *Biochem. J.* **261**, 951-958
- Conlin, C. A., and Miller, C. G. (1995) *Methods Enzymol.* **248**, 567-579
- Monnet, V. (1995) *Methods Enzymol.* **248**, 579-592
- Isaya, G., Kalousek, F., and Rosenberg, L. E. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89**, 8317-8321
- Isaya, G., Miklos, D., and Rollins, R. A. (1994) *Mol. Cell Biol.* **14**, 5603-5616
- Buchler, M., Tislijar, U., and Wolf, D. H. (1994) *Eur. J. Biochem.* **219**, 627-639
- Rawlings, N. D., and Barrett, A. J. (1995) *Methods Enzymol.* **248**, 183-228
- Rawlings, N. D., and Barrett, A. J. (1993) *Biochem. J.* **290**, 205-218
- Wünsch, E., and Heidrich, H.-G. (1963) *Hoppe-Seyler's Z. Physiol. Chem.* **333**, 149-151
- Harper, E., and Gross, J. (1970) *Biochim. Biophys. Acta* **198**, 286-292
- Cauncho, A. C., Ramalho-Pinto, F. J., and Greene, L. J. (1972) *J. Neurochem.* **19**, 37-49
- Orlowski, M., Michaud, C., and Chu, T. G. (1983) *Eur. J. Biochem.* **135**, 81-88
- Barrett, A. J., and Brown, M. A. (1990) *Biochem. J.* **271**, 701-706
- McKie, N., Dando, P. M., Rawlings, N. D., and Barrett, A. J. (1993) *Biochem. J.* **295**, 57-60
- Pierotti, A., Dong, K. W., Glucksman, M. J., Orlowski, M., and Roberts, J. L. (1994) *Biochemistry* **33**, 622
- Kato, A., Sugiura, N., Hagiwara, H., and Hirose, S. (1994) *Eur. J. Biochem.* **221**, 159-165
- Thompson, A., Huber, G., and Malherbe, P. (1995) *Biochem. Biophys. Res. Commun.* **213**, 65-73
- Heidrich, H.-G., Prokopenko, D., and Hanning, K. (1969) *Hoppe-Seyler's Z. Physiol. Chem.* **350**, 1430-1436
- Orlowski, M., and Barrett, A. J. (1990) *FEBS Lett.* **264**, 84-86
- Serizawa, A., Dando, P. M., and Barrett, A. J. (1995) *J. Biol. Chem.* **270**, 2092-2098
- Hagiwara, H., Sugiura, N., Wakita, K., and Hirose, S. (1989) *Eur. J. Biochem.* **185**, 405-410
- Kiron, M. A. R., and Soffer, R. L. (1989) *J. Biol. Chem.* **264**, 4138-4142
- Sugiura, N., Hagiwara, H., and Hirose, S. (1992) *J. Biol. Chem.* **267**, 18067-18072
- Kawabata, S., and Davie, E. W. (1992) *J. Biol. Chem.* **267**, 10331-10336
- Kawabata, S., Nakagawa, K., Muta, T., Iwanaga, S., and Davie, E. W. (1993) *J. Biol. Chem.* **268**, 12498-12503
- Checler, F., Vincent, J. P., and Kitabgi, P. (1983) *J. Neurochem.* **41**, 375-384
- Checler, F., Emson, P. C., Vincent, J. P., and Kitabgi, P. (1984) *J. Neurochem.* **43**, 1295-1301
- Checler, F., Vincent, J.-P., and Kitabgi, P. (1986) *J. Biol. Chem.* **261**, 11274-11281
- Dauch, P., Vincent, J.-P., and Checler, F. (1995) *J. Biol. Chem.* **270**, 27265-27271
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467
- Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349-383
- von Heijne, G., Steppuhn, J., and Herrmann, R. G. (1989) *Eur. J. Biochem.* **190**, 535-545
- Davies, M. S., Henney, A., Ward, W. H., and Craig, R. K. (1986) *Gene (Amst.)* **45**, 183-191
- Williams, T., Admon, A., Luscher, B., and Tjian, R. (1988) *Genes Dev.* **2**, 1557-1569
- Imagawa, M., Chiu, R., and Karin, M. (1987) *Cell* **51**, 251-260
- Hyman, S. E., Comb, M., Pearlberg, J., and Goodman, H. M. (1989) *Mol. Cell Biol.* **9**, 321-324
- Kadonaga, J. T., Jones, K. A., and Tjian, R. (1997) *Trends Biochem. Sci.* **11**, 20-23 (abstr.)
- Luscher, B., and Eisenman, R. N. (1990) *Genes Dev.* **4**, 2235-2241
- Ko, L. J., and Engel, J. D. (1993) *Mol. Cell Biol.* **13**, 4011-4022
- Mizuguchi, G., Kanei-Ishii, C., Takahashi, T., Yasukawa, T., Nagase, T., Horikoshi, M., Yamamoto, T., and Ishii, S. (1995) *J. Biol. Chem.* **270**, 9384-9389
- Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125-8148
- Hardt, F. U., Pfanner, N., Nicholson, D. W., and Neupert, W. (1989) *Biochim. Biophys. Acta* **988**, 1-45
- Danpure, C. J. (1995) *Trends Cell Biol.* **5**, 230-238
- Natsoulis, G., Hilger, F., and Fink, G. R. (1986) *Cell* **46**, 235-243
- Chatton, B., Walter, P., Ebel, J.-P., Lacroute, F., and Fasiolo, F. (1988) *J. Biol. Chem.* **263**, 52-57
- Rosenberg, E., Russo, C., Kiron, M. A. R., and Soffer, R. L. (1988) *Biochem. Biophys. Res. Commun.* **156**, 1230-1236
- Sugiura, N., Hagiwara, H., Ito, T., and Hirose, S. (1991) in *Peptide Regulation of Cardiovascular Function* (Imura, H., Matsuo, H., and Masaki, T., eds) pp. 157-166, Academic Press, New York
- Conlin, C. A., Trun, N. J., Silhavy, T. J., and Miller, C. G. (1992) *J. Bacteriol.* **174**, 5881-5887
- Henrich, B., Becker, S., Schroeder, U., and Plapp, R. (1993) *J. Bacteriol.* **175**, 7290-7300
- Serizawa, A., Dando, P. M., and Barrett, A. J. (1997) in *Proteolysis in Cell Functions* (Hopsu-Havu, V. K., Järvinen, M., and Kirschke, H. eds) IOS Press, Amsterdam, in press

<sup>2</sup> A. Kato, N. Sugiura, and S. Hirose, unpublished observation.



National  
Library  
of Medicine 

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM [Book](#)

Search [PubMed](#) for  [Go](#) [Clear](#)

Limits

Preview/Index

History

Clipboard

Details

[About Entrez](#)

[Text Version](#)

Entrez PubMed

[Overview](#)

[Help | FAQ](#)

[Tutorial](#)

[New/Noteworthy](#)

[E-Utilities](#)

PubMed Services

[Journals Database](#)

[MeSH Browser](#)

[Single Citation Matcher](#)

[Batch Citation Matcher](#)

[Clinical Queries](#)

[LinkOut](#)

[Cubby](#)

Related Resources

[Order Documents](#)

[NLM Gateway](#)

[TOXNET](#)

[Consumer Health](#)

[Clinical Alerts](#)

[ClinicalTrials.gov](#)

[PubMed Central](#)

[Privacy Policy](#)

•  1: [Endocr Rev 2002 Oct;23\(5\):647-64](#)

[Related Articles, Links](#)

[Full text article at  
cdrv.endojournals.org](#)

## Soluble metalloendopeptidases and neuroendocrine signaling.

Shrimpton CN, Smith AI, Lew RA.

Baker Medical Research Institute (C.N.S., A.I.S., R.A.L.), Melbourne, Australia 8008.

Peptidases play a vital and often highly specific role in the physiological and pathological generation and termination of peptide hormone signals. The thermolysin-like family of metalloendopeptidases involved in the extracellular processing of neuroendocrine and cardiovascular peptides are of particular significance, reflecting both their specificity for particular peptide substrates and their utility as therapeutic targets. Although the functions of the membrane-bound members of this family, such as angiotensin-converting enzyme and neutral endopeptidase, are well established, a role for the predominantly soluble family members in peptide metabolism is only just emerging. This review will focus on the biochemistry, cell biology, and physiology of the soluble metalloendopeptidases EC 3.4.24.15 (thimet oligopeptidase) and EC 3.4.24.16 (neurolysin), as well as presenting evidence that both peptidases play an important role in such diverse functions as reproduction, nociception, and cardiovascular homeostasis.

PMID: 12372844 [PubMed - in process]

[Display](#) [Abstract](#) [Sort](#) [Save](#) [Text](#) [Clip Add](#) [Order](#)

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Freedom of Information Act](#) | [Disclaimer](#)

i686-pc-linux-gnu Oct 21 2002 11:24:19

# The Sequence of the Human Genome

J. Craig Venter,<sup>1\*</sup> Mark D. Adams,<sup>1</sup> Eugene W. Myers,<sup>1</sup> Peter W. Li,<sup>1</sup> Richard J. Mural,<sup>1</sup> Granger G. Sutton,<sup>1</sup> Hamilton O. Smith,<sup>1</sup> Mark Yandell,<sup>1</sup> Cheryl A. Evans,<sup>1</sup> Robert A. Holt,<sup>1</sup> Jeannine D. Gocayne,<sup>1</sup> Peter Amanatides,<sup>1</sup> Richard M. Ballew,<sup>1</sup> Daniel H. Huson,<sup>1</sup> Jennifer Russo Wortman,<sup>1</sup> Qing Zhang,<sup>1</sup> Chinnappa D. Kodira,<sup>1</sup> Xiangqun H. Zheng,<sup>1</sup> Lin Chen,<sup>1</sup> Marian Skupski,<sup>1</sup> Gangadharan Subramanian,<sup>1</sup> Paul D. Thomas,<sup>1</sup> Jinghui Zhang,<sup>1</sup> George L. Gabor Miklos,<sup>2</sup> Catherine Nelson,<sup>3</sup> Samuel Broder,<sup>1</sup> Andrew G. Clark,<sup>4</sup> Joe Nadeau,<sup>5</sup> Victor A. McKusick,<sup>6</sup> Norton Zinder,<sup>7</sup> Arnold J. Levine,<sup>7</sup> Richard J. Roberts,<sup>8</sup> Mel Simon,<sup>9</sup> Carolyn Slayman,<sup>10</sup> Michael Hunkapiller,<sup>11</sup> Randall Bolanos,<sup>1</sup> Arthur Delcher,<sup>1</sup> Ian Dew,<sup>1</sup> Daniel Fasulo,<sup>1</sup> Michael Flanigan,<sup>1</sup> Liliana Florea,<sup>1</sup> Aaron Halpern,<sup>1</sup> Sridhar Hannenhalli,<sup>1</sup> Saul Kravitz,<sup>1</sup> Samuel Levy,<sup>1</sup> Clark Moberly,<sup>1</sup> Knut Reinert,<sup>1</sup> Karin Remington,<sup>1</sup> Jane Abu-Threideh,<sup>1</sup> Ellen Beasley,<sup>1</sup> Kendra Biddick,<sup>1</sup> Vivien Bonazzi,<sup>1</sup> Rhonda Brandon,<sup>1</sup> Michele Cargill,<sup>1</sup> Ishwar ChandramouliSwaran,<sup>1</sup> Rosane Charlab,<sup>1</sup> Kabir Chaturvedi,<sup>1</sup> Zuoming Deng,<sup>1</sup> Valentina Di Francesco,<sup>1</sup> Patrick Dunn,<sup>1</sup> Karen Eilbeck,<sup>1</sup> Carlos Evangelista,<sup>1</sup> Andrei E. Gabrielian,<sup>1</sup> Weinu Gan,<sup>1</sup> Wangmao Ge,<sup>1</sup> Fangcheng Gong,<sup>1</sup> Zhiping Gu,<sup>1</sup> Ping Guan,<sup>1</sup> Thomas J. Heiman,<sup>1</sup> Maureen E. Higgins,<sup>1</sup> Rui-Ru Ji,<sup>1</sup> Zhaoxi Ke,<sup>1</sup> Karen A. Ketchum,<sup>1</sup> Zhongwu Lai,<sup>1</sup> Yiding Lei,<sup>1</sup> Zhenya Li,<sup>1</sup> Jiayin Li,<sup>1</sup> Yong Liang,<sup>1</sup> Xiaoying Lin,<sup>1</sup> Fu Lu,<sup>1</sup> Gennady V. Merkulov,<sup>1</sup> Natalia Milshina,<sup>1</sup> Helen M. Moore,<sup>1</sup> Ashwinikumar K Naik,<sup>1</sup> Vaibhav A. Narayan,<sup>1</sup> Beena Neelam,<sup>1</sup> Deborah Nusskern,<sup>1</sup> Douglas B. Rusch,<sup>1</sup> Steven Salzberg,<sup>12</sup> Wei Shao,<sup>1</sup> Bixiong Shue,<sup>1</sup> Jingtao Sun,<sup>1</sup> Zhen Yuan Wang,<sup>1</sup> Aihui Wang,<sup>1</sup> Xin Wang,<sup>1</sup> Jian Wang,<sup>1</sup> Ming-Hui Wei,<sup>1</sup> Ron Wides,<sup>13</sup> Chunlin Xiao,<sup>1</sup> Chunhua Yan,<sup>1</sup> Alison Yao,<sup>1</sup> Jane Ye,<sup>1</sup> Ming Zhan,<sup>1</sup> Weiqing Zhang,<sup>1</sup> Hongyu Zhang,<sup>1</sup> Qi Zhao,<sup>1</sup> Liansheng Zheng,<sup>1</sup> Fei Zhong,<sup>1</sup> Wenyan Zhong,<sup>1</sup> Shaooping C. Zhu,<sup>1</sup> Shaying Zhao,<sup>12</sup> Dennis Gilbert,<sup>1</sup> Suzanna Baumhueter,<sup>1</sup> Gene Spier,<sup>1</sup> Christine Carter,<sup>1</sup> Anibal Cravchik,<sup>1</sup> Trevor Woodage,<sup>1</sup> Feroze Ali,<sup>1</sup> Huijin An,<sup>1</sup> Aderonke Awe,<sup>1</sup> Danita Baldwin,<sup>1</sup> Holly Baden,<sup>1</sup> Mary Barnstead,<sup>1</sup> Ian Barrow,<sup>1</sup> Karen Beeson,<sup>1</sup> Dana Busam,<sup>1</sup> Amy Carver,<sup>1</sup> Angela Center,<sup>1</sup> Ming Lai Cheng,<sup>1</sup> Liz Curry,<sup>1</sup> Steve Danaher,<sup>1</sup> Lionel Davenport,<sup>1</sup> Raymond Desilets,<sup>1</sup> Susanne Dietz,<sup>1</sup> Kristina Dodson,<sup>1</sup> Lisa Doup,<sup>1</sup> Steven Ferriera,<sup>1</sup> Neha Garg,<sup>1</sup> Andres Gluecksmann,<sup>1</sup> Brit Hart,<sup>1</sup> Jason Haynes,<sup>1</sup> Charles Haynes,<sup>1</sup> Cheryl Heiner,<sup>1</sup> Suzanne Hladun,<sup>1</sup> Damon Hostin,<sup>1</sup> Jarrett Houck,<sup>1</sup> Timothy Howland,<sup>1</sup> Chinyere Ibegwam,<sup>1</sup> Jeffery Johnson,<sup>1</sup> Francis Kalush,<sup>1</sup> Lesley Kline,<sup>1</sup> Shashi Koduru,<sup>1</sup> Amy Love,<sup>1</sup> Felecia Mann,<sup>1</sup> David May,<sup>1</sup> Steven McCawley,<sup>1</sup> Tina McIntosh,<sup>1</sup> Ivy McMullen,<sup>1</sup> Mee Moy,<sup>1</sup> Linda Moy,<sup>1</sup> Brian Murphy,<sup>1</sup> Keith Nelson,<sup>1</sup> Cynthia Pfannkoch,<sup>1</sup> Eric Pratts,<sup>1</sup> Vinita Puri,<sup>1</sup> Hina Qureshi,<sup>1</sup> Matthew Reardon,<sup>1</sup> Robert Rodriguez,<sup>1</sup> Yu-Hui Rogers,<sup>1</sup> Deanna Romblad,<sup>1</sup> Bob Ruhfel,<sup>1</sup> Richard Scott,<sup>1</sup> Cynthia Sitter,<sup>1</sup> Michelle Smallwood,<sup>1</sup> Erin Stewart,<sup>1</sup> Renee Strong,<sup>1</sup> Ellen Suh,<sup>1</sup> Reginald Thomas,<sup>1</sup> Ni Ni Tint,<sup>1</sup> Sukyee Tse,<sup>1</sup> Claire Vech,<sup>1</sup> Gary Wang,<sup>1</sup> Jeremy Wetter,<sup>1</sup> Sherita Williams,<sup>1</sup> Monica Williams,<sup>1</sup> Sandra Windsor,<sup>1</sup> Emily Winn-Deen,<sup>1</sup> Keriellen Wolfe,<sup>1</sup> Jayshree Zaveri,<sup>1</sup> Karena Zaveri,<sup>1</sup> Josep F. Abril,<sup>14</sup> Roderic Guigó,<sup>14</sup> Michael J. Campbell,<sup>1</sup> Kimmen V. Sjolander,<sup>1</sup> Brian Karlak,<sup>1</sup> Anish Kejariwal,<sup>1</sup> Huaiyu Mi,<sup>1</sup> Betty Lazareva,<sup>1</sup> Thomas Hatton,<sup>1</sup> Apurva Narechania,<sup>1</sup> Karen Diemer,<sup>1</sup> Anushya Muruganujan,<sup>1</sup> Nan Guo,<sup>1</sup> Shinji Sato,<sup>1</sup> Vineet Bafna,<sup>1</sup> Sorin Istrail,<sup>1</sup> Ross Lippert,<sup>1</sup> Russell Schwartz,<sup>1</sup> Brian Walenz,<sup>1</sup> Shibu Yooseph,<sup>1</sup> David Allen,<sup>1</sup> Anand Basu,<sup>1</sup> James Baxendale,<sup>1</sup> Louis Blick,<sup>1</sup> Marcelo Caminha,<sup>1</sup> John Carnes-Stine,<sup>1</sup> Parris Caulk,<sup>1</sup> Yen-Hui Chiang,<sup>1</sup> My Coyne,<sup>1</sup> Carl Dahlke,<sup>1</sup> Anne Deslattes Mays,<sup>1</sup> Maria Dombroski,<sup>1</sup> Michael Donnelly,<sup>1</sup> Dale Ely,<sup>1</sup> Shiva Esparham,<sup>1</sup> Carl Fosler,<sup>1</sup> Harold Gire,<sup>1</sup> Stephen Glatowski,<sup>1</sup> Kenneth Glasser,<sup>1</sup> Anna Glodek,<sup>1</sup> Mark Gorokhov,<sup>1</sup> Ken Graham,<sup>1</sup> Barry Gropman,<sup>1</sup> Michael Harris,<sup>1</sup> Jeremy Heil,<sup>1</sup> Scott Henderson,<sup>1</sup> Jeffrey Hoover,<sup>1</sup> Donald Jennings,<sup>1</sup> Catherine Jordan,<sup>1</sup> James Jordan,<sup>1</sup> John Kasha,<sup>1</sup> Leonid Kagan,<sup>1</sup> Cheryl Kraft,<sup>1</sup> Alexander Levitsky,<sup>1</sup> Mark Lewis,<sup>1</sup> Xiangjun Liu,<sup>1</sup> John Lopez,<sup>1</sup> Daniel Ma,<sup>1</sup> William Majoros,<sup>1</sup> Joe McDaniel,<sup>1</sup> Sean Murphy,<sup>1</sup> Matthew Newman,<sup>1</sup> Trung Nguyen,<sup>1</sup> Ngoc Nguyen,<sup>1</sup> Marc Nodell,<sup>1</sup> Sue Pan,<sup>1</sup> Jim Peck,<sup>1</sup> Marshall Peterson,<sup>1</sup> William Rowe,<sup>1</sup> Robert Sanders,<sup>1</sup> John Scott,<sup>1</sup> Michael Simpson,<sup>1</sup> Thomas Smith,<sup>1</sup> Arlan Sprague,<sup>1</sup> Timothy Stockwell,<sup>1</sup> Russell Turner,<sup>1</sup> Eli Venter,<sup>1</sup> Mei Wang,<sup>1</sup> Meiyuan Wen,<sup>1</sup> David Wu,<sup>1</sup> Mitchell Wu,<sup>1</sup> Ashley Xia,<sup>1</sup> Ali Zandieh,<sup>1</sup> Xiaohong Zhu<sup>1</sup>

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward un-

derstanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (1). In subsequent years, the idea met with mixed reactions in the scientific community (2). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, \$3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of

DNA using chain-terminating nucleotide analogs (3). In the same year, the first human gene was isolated and sequenced (4). In 1986, Hood and co-workers (5) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (6). From early sequencing of human genomic regions (7), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (8), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (9). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (10).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (11). When considering methods for sequencing the smallpox virus genome in 1991 (12), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (13). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (14, 15).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (16) of an approach to simulta-

<sup>1</sup>Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. <sup>2</sup>GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. <sup>3</sup>Berkeley Drosophila Genome Project, University of California, Berkeley, CA 94720, USA. <sup>4</sup>Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. <sup>5</sup>Department of Genetics, Case Western Reserve University School of Medicine, BBR-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. <sup>6</sup>Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287-4922, USA. <sup>7</sup>Rockefeller University, 1230 York Avenue, New York, NY 10021-6399, USA. <sup>8</sup>New England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA. <sup>9</sup>Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. <sup>10</sup>Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520-8000, USA. <sup>11</sup>Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. <sup>12</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. <sup>13</sup>Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. <sup>14</sup>Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

\*To whom correspondence should be addressed. E-mail: humangenome@celera.com

## THE HUMAN GENOME

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (17, 18). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (19).

In 1997, Weber and Myers (20) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (21). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (22), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (23). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (24). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (25). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (26–28). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (29), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold

coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (30) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC1)) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

- 1 Sources of DNA and Sequencing Methods
- 2 Genome Assembly Strategy and Characterization
- 3 Gene Prediction and Annotation
- 4 Genome Structure
- 5 Genome Evolution
- 6 A Genome-Wide Examination of Sequence Variations
- 7 An Overview of the Predicted Protein-Coding Genes in the Human Genome
- 8 Conclusions

### 1 Sources of DNA and Sequencing Methods

**Summary.** This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genomic sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (31) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds. Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (32).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of semen, collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at [www.sciencemag.org/cgi/content/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/291/5507/1304/DC1)). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

### 1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (33).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (34).

Current sequencing protocols are based on

the dideoxy sequencing method (35), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (36).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis

through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

### 1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (37). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (26). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

### 1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

No. of sequencing reads	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)		0	0	0.52	0.52	
	A	2.20	1.40	0.01	3.61	
	B	0.16	1.17	0	0.32	
	C	0.18	0.20	0	0.37	
	D	0	0.28	0	0.28	
	F	2.54	2.04	0.53	5.11	
	Total					
Fold clone coverage		0	0	18.39	18.39	
	A	2.96	11.26	0.44	14.67	
	B	0.22	1.33	0	1.54	
	C	0.24	1.58	0	1.82	
	D	0	2.26	0	2.26	
	F	3.42	16.43	18.84	38.68	
	Total					
Insert size* (mean)		Average	1,951 bp	10,800 bp	50,715 bp	
Insert size* (SD)		Average	6.10%	8.10%	14.90%	
% Mates†		Average	74.50	80.80	75.60	

\*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

## THE HUMAN GENOME

name, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (26). By collecting data for the

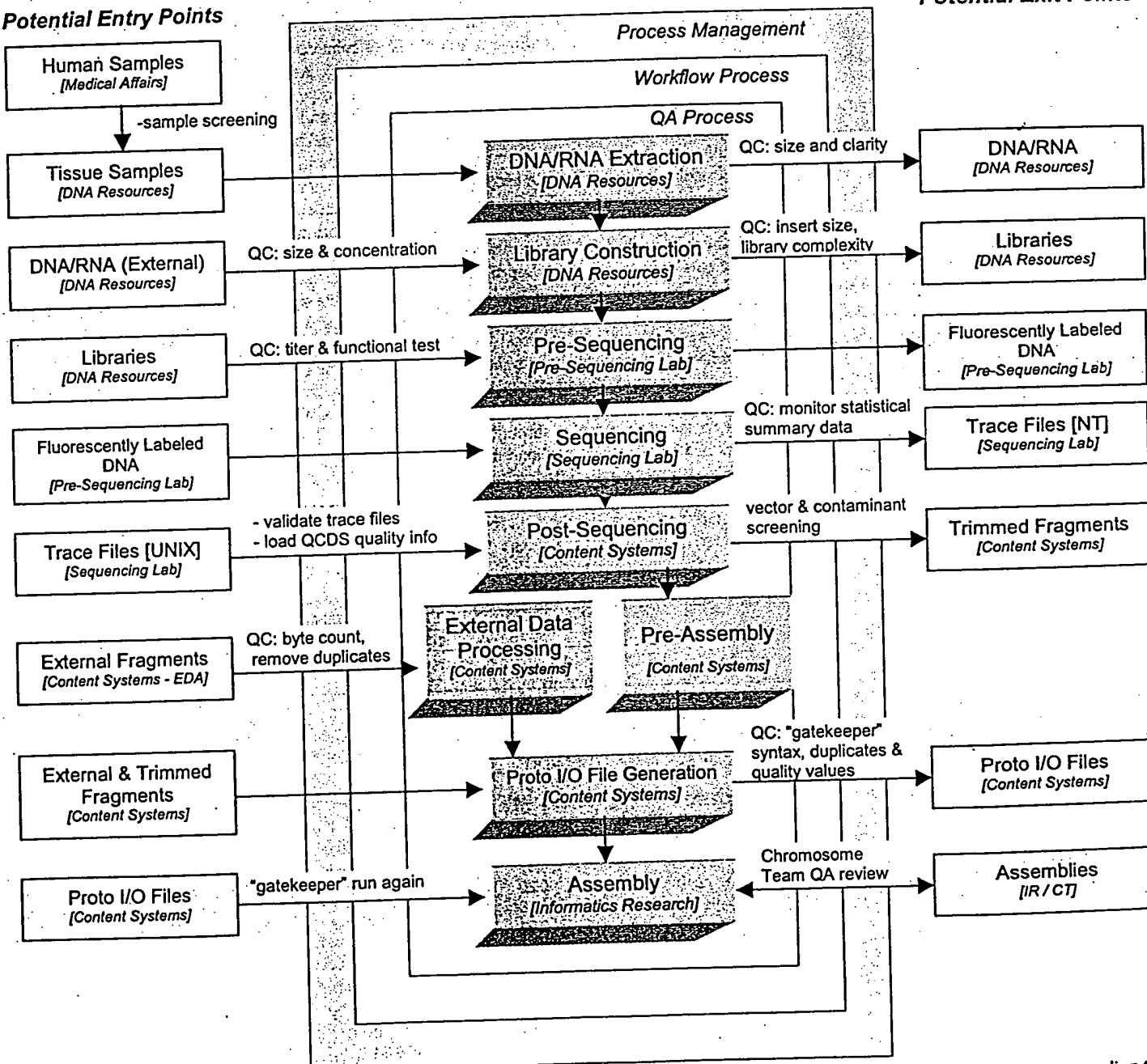
entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

### 2 Genome Assembly Strategy and Characterization

**Summary.** We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an indepen-

dent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process.

#### Potential Entry Points



**Fig. 2.** Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange

samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public genome sequence, which was reconstructed largely by an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff" and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

## 2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (HGP) and is primarily derived from BAC clones (30). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC

sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (38), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Sequences division of GenBank (39), filtered at 200 bp at 98%; and (iii) the non-redundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (18).

## 2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (40). Furthermore, BAC location

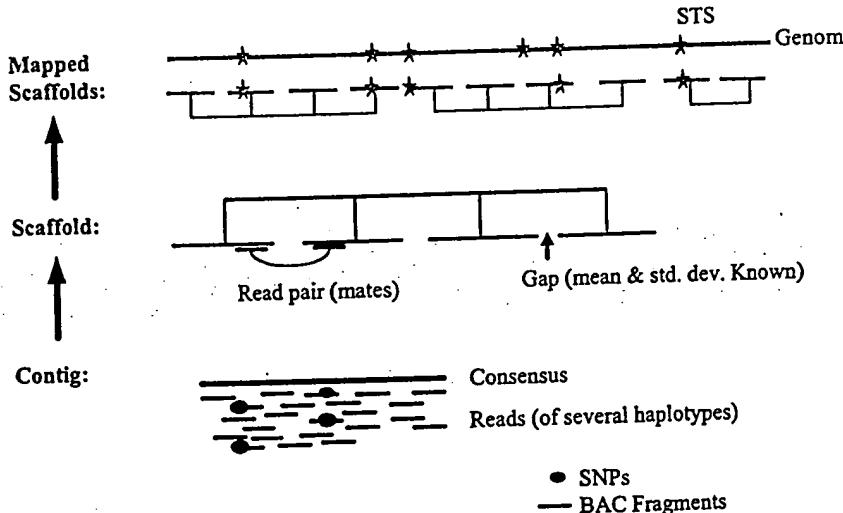


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

# THE HUMAN GENOME

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that

at least 2.2% of the BACs contained sequence data that were not part of the given BAC (41), possibly as a result of sample-tracking errors

**Table 2.** GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
Washington University, USA	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
Baylor College of Medicine, USA	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
Others*	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

\*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research; Gene Bank; The University of Oklahoma; University of Texas Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Research; Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96X coverage of the genome.

(see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in the reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5X Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile these scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and curated the scaffold tiling of the components to further increase its accuracy. For the final CS assembly, all but the partitioning was ignored and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data in the shredded, faux reads of the partitioned, relevant bactig data.

### 2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome report in detail in (28).

The WGA assembler consists of a pipeline composed of five principal stages: Screen Overlapper, Unitigger, Scaffold, and Resolver, respectively. The Screen Overlapper finds and marks all microsatellite repeats with less than a 6-bp element, and screens out known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Mismatches get searched for overlaps, while screened regions do not get searched, but can be part of an overlap that involves unscanned matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in- $10^{17}$  event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (42). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6 $\times$  simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all

singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in  $10^{10}$ , assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more

aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than  $10^{-7}$  based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (43). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

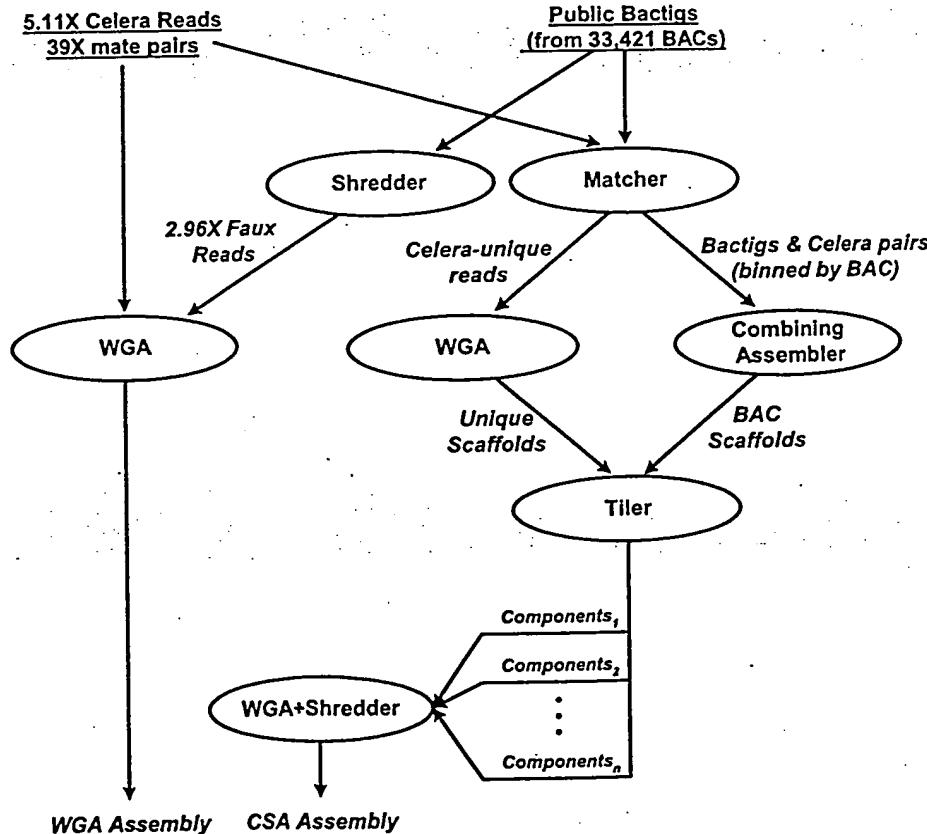


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value-weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence-construction subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would

have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the dis-

tribution of each was essentially exponential. More than 50% of all gaps were less than 500 bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

## 2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
<i>Compartmentalized shotgun assembly</i>					
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤ 1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
<i>Whole-genome assembly</i>					
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤ 1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11 $\times$  redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5 $\times$  Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of

assembly took place, but not enough Celera data were matched to truly assemble the 0.5 $\times$  to 1 $\times$  data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1 $\times$  light-shotgun of BACs will not yield good assembly of BAC regions; at least 3 $\times$  light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (18) and sequence tagged site (STS) markers (44) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2 $\times$  shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry.

Chimeric or contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs >30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

## 2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were

not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely  $>1$  kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

## 2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (45). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those  $>9$  Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (46) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds.

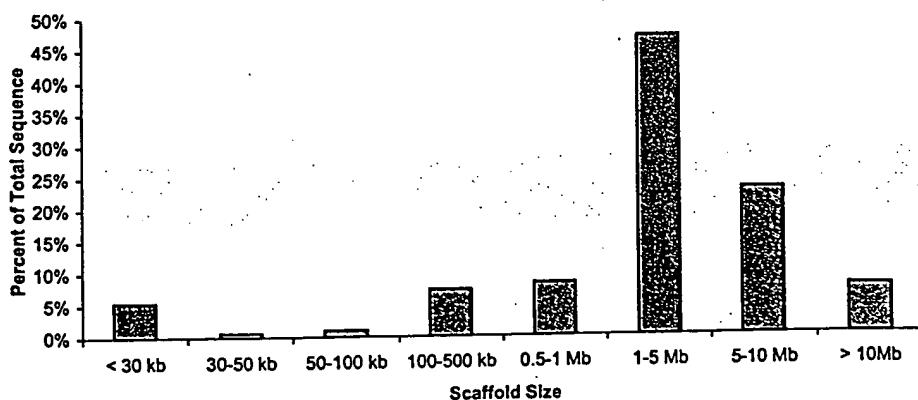


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the genome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (47). This is consistent with the bacterial chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have forced certain regions of the assembly by arranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

## 2.7 Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

**Completeness.** Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (48, 49); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (50, 51).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (48, 49). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (52). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (51) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (53) and BLAST (54) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (55a) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

**Correctness.** Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the assembly.

**Table 4. Summary of scaffold mapping.** Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

## THE HUMAN GENOME

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean  $\pm 3$  standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "mis-separated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean  $\pm$  the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (48) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for

those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and mis-separated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39 $\times$ , meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds  $>30$  kbp in length, less than 1% of the Celera assembly was in regions of less than 3 $\times$  clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of

5 September 2000 (30, 55b). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

**Table 5.** Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number

of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

Library type	Library no.	Chromosome 21					Genome			
		Mean insert size (bp)	SD (bp)	SD/mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768	2.7			
						(mean = 2.7)				

## THE HUMAN GENOME

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

### 3 Gene Prediction and Annotation

**Summary.** To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

#### 3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alterna-

tive splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (56). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density-based extrapolations, have not reduced this variance. The highest recent report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (57). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (58); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (59); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can

be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (60). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (61)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (62). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g., protein, EST, and so forth). The results were used to group the matches into bins of related sequences that may define a gene and identif

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.\*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

\*Data for individual chromosomes can be found in Web fig. 3 on Science Online at [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). †Mates are misseparated if their distance is >3 SD from the mean library size.

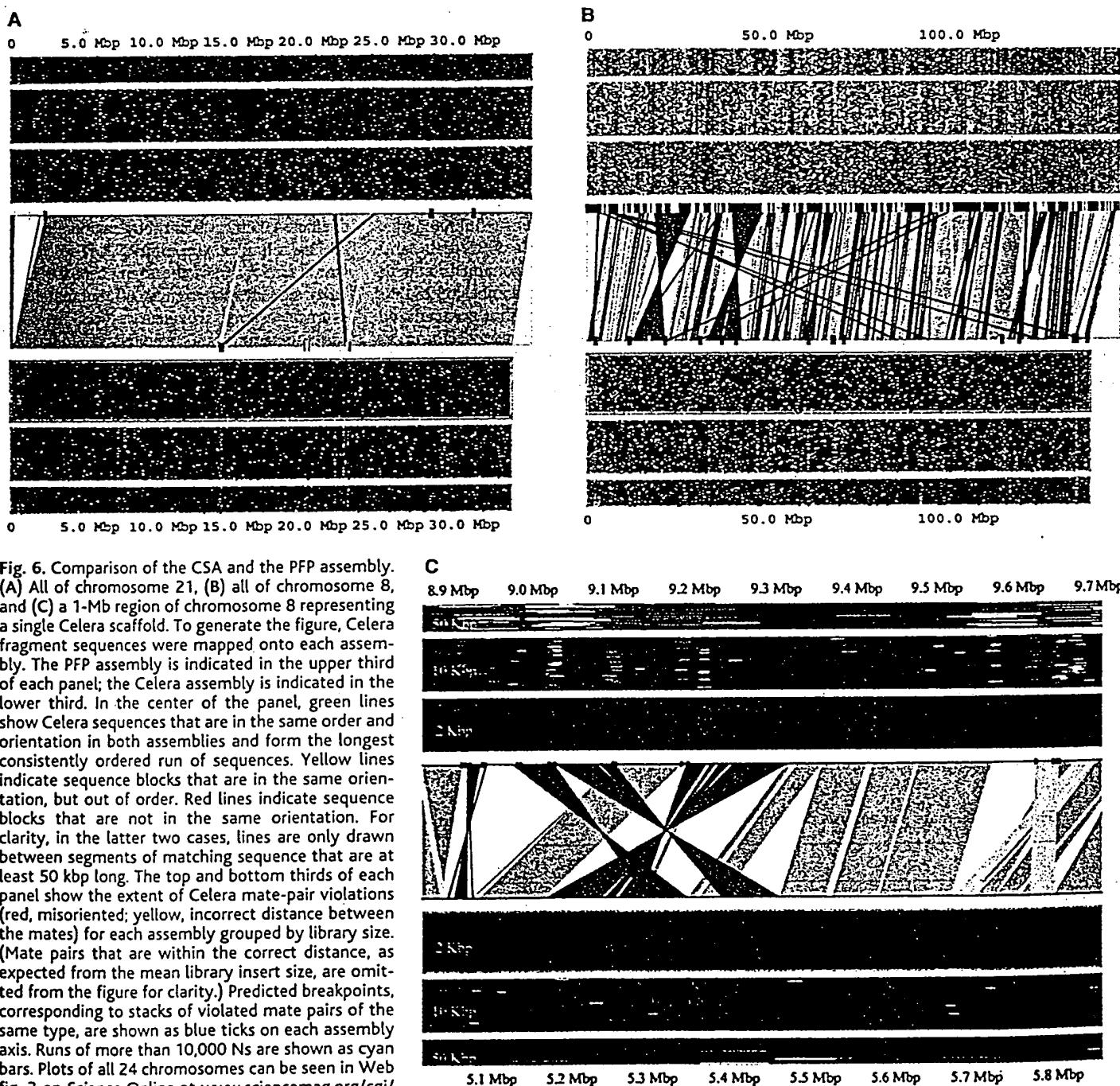
gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes

being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (63) alignment of the RefSeq transcript to

the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto



**Fig. 6.** Comparison of the CSA and the PFP assembly. (A) All of chromosome 21, (B) all of chromosome 8, and (C) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on Science Online at [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC1).

## THE HUMAN GENOME

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs

and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the hu-

man genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

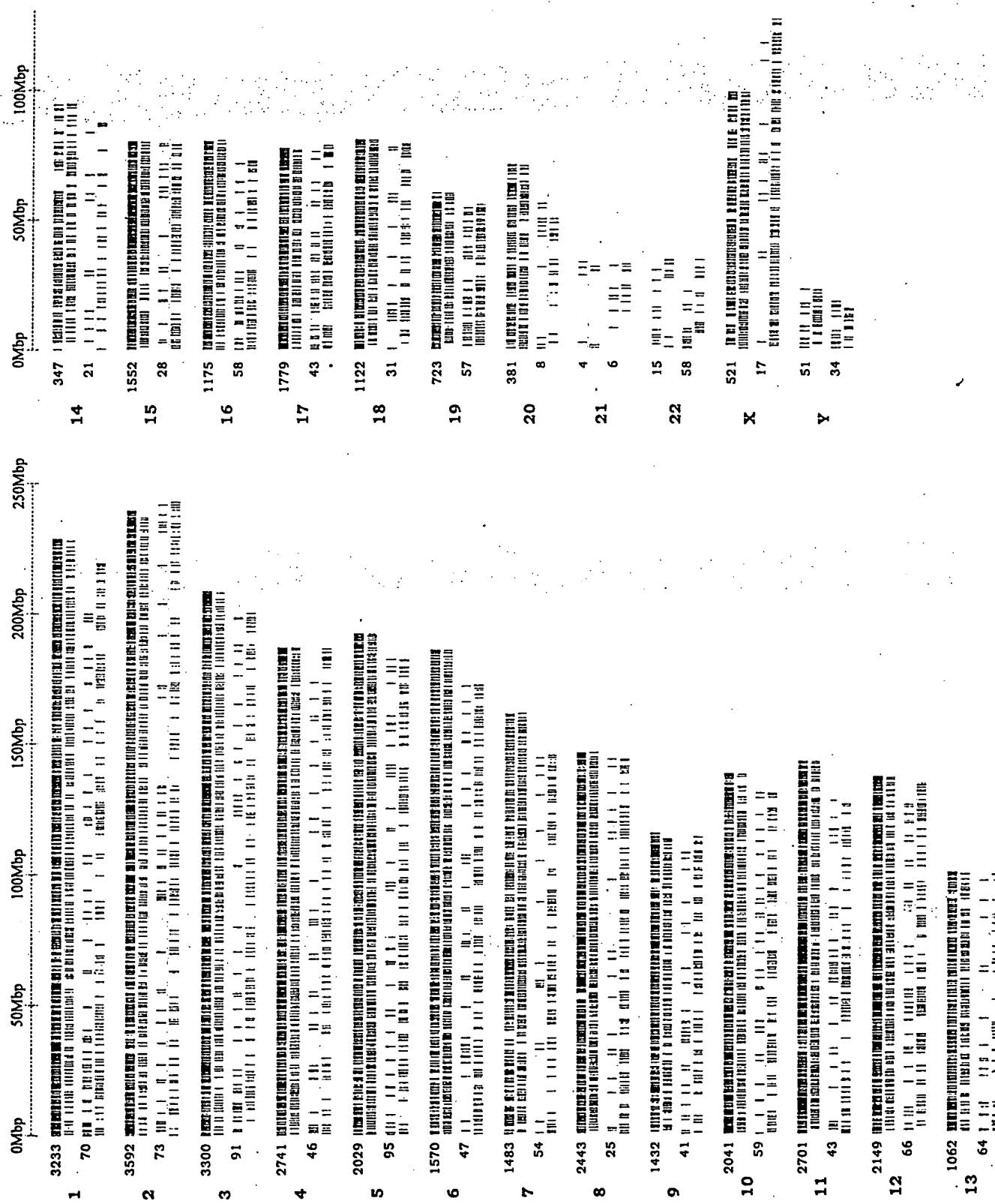


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's

assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within  $\pm 10$  bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5' and 3' untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be  $>0.66$  or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits ( $\pm 10$  bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and

those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3' and 5' untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (63)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

### 3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertently result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

### 3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which  $\sim 76,410$  were nonredundant (non-overlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to  $\sim 23,000$ . Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3' UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of

**Table 7.** Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; *P* < 0.001).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

\*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

### 3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that sup-

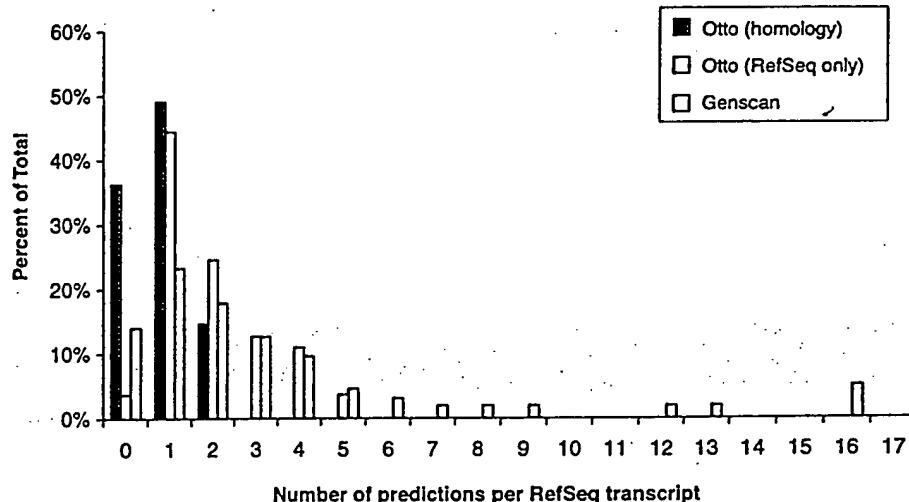
port the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

### 4 Genome Structure

**Summary.** This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

### 4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (64). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (65). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (66). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.



**Fig. 8.** Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

**Table 8.** Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	17,968†	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	21,350	8,619	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript		Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00
		De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42
*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.										

\*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

## THE HUMAN GENOME

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (67). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (68). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (69). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (70). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (69). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we

found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (71) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

### 4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis.

In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (72). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

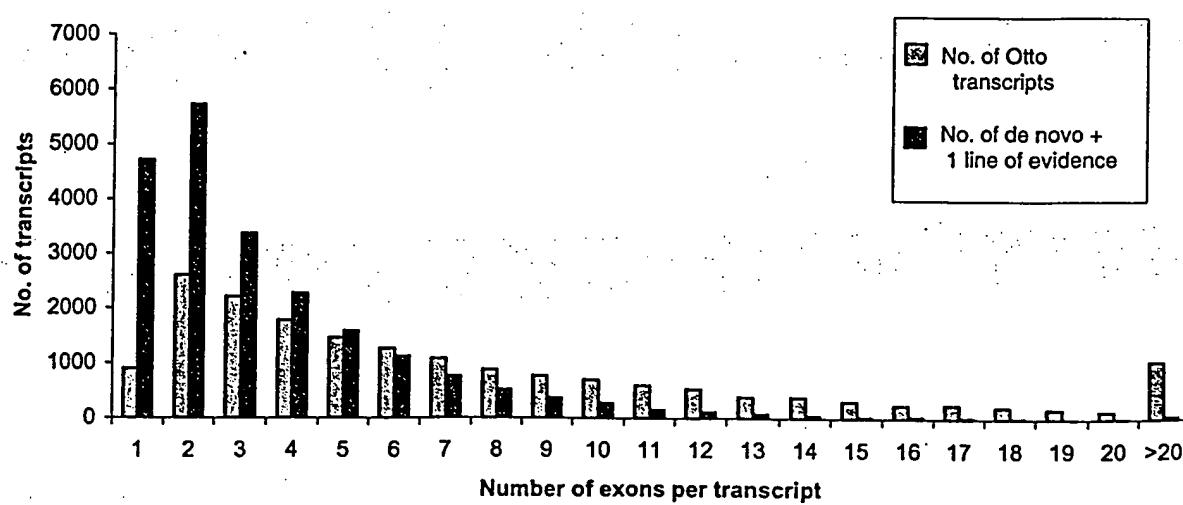
We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (73). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window.

Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43-48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

\*The predictions were based on Bernardi's definitions (70) of the isochore structure of the human genome.

Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.



examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Etude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate of variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

#### 4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (74). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5' end of the transcript (75, 76). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (77) and has been shown to be important during gene imprinting (78) and tissue-specific gene expression (79).

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (74, 80) and an estimate of 499 CpG islands on human chromosome 22 (81). Larsen *et al.* (76) and Gardiner-Garden and Frommer (75) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed

versus expected frequency of CG dinucleotide  $\geq 0.6$ .

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al.* (76). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a

CpG island. This is comparable to ratios reported by others (82). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

#### 4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (83). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

#### 5 Genome Evolution

**Summary.** The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

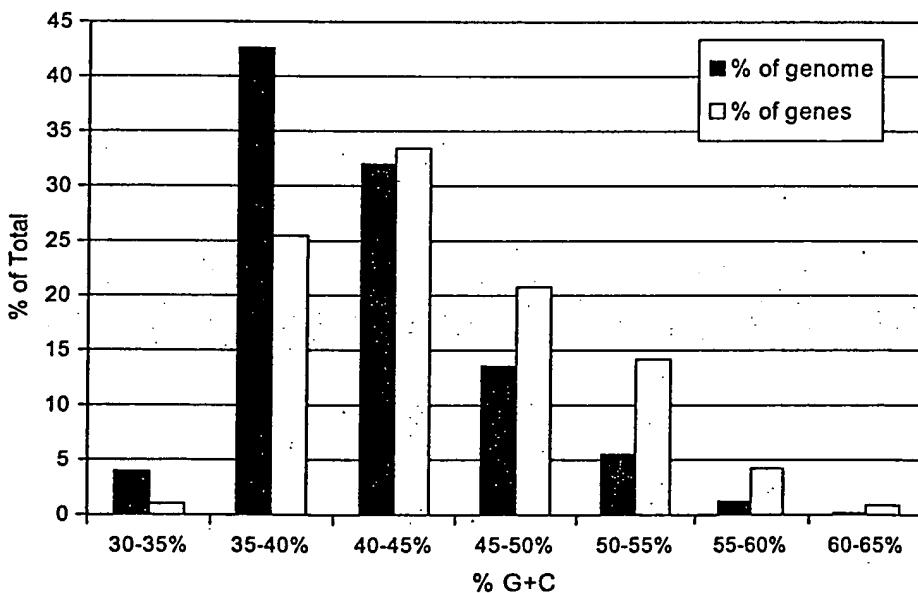


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

## THE HUMAN GENOME

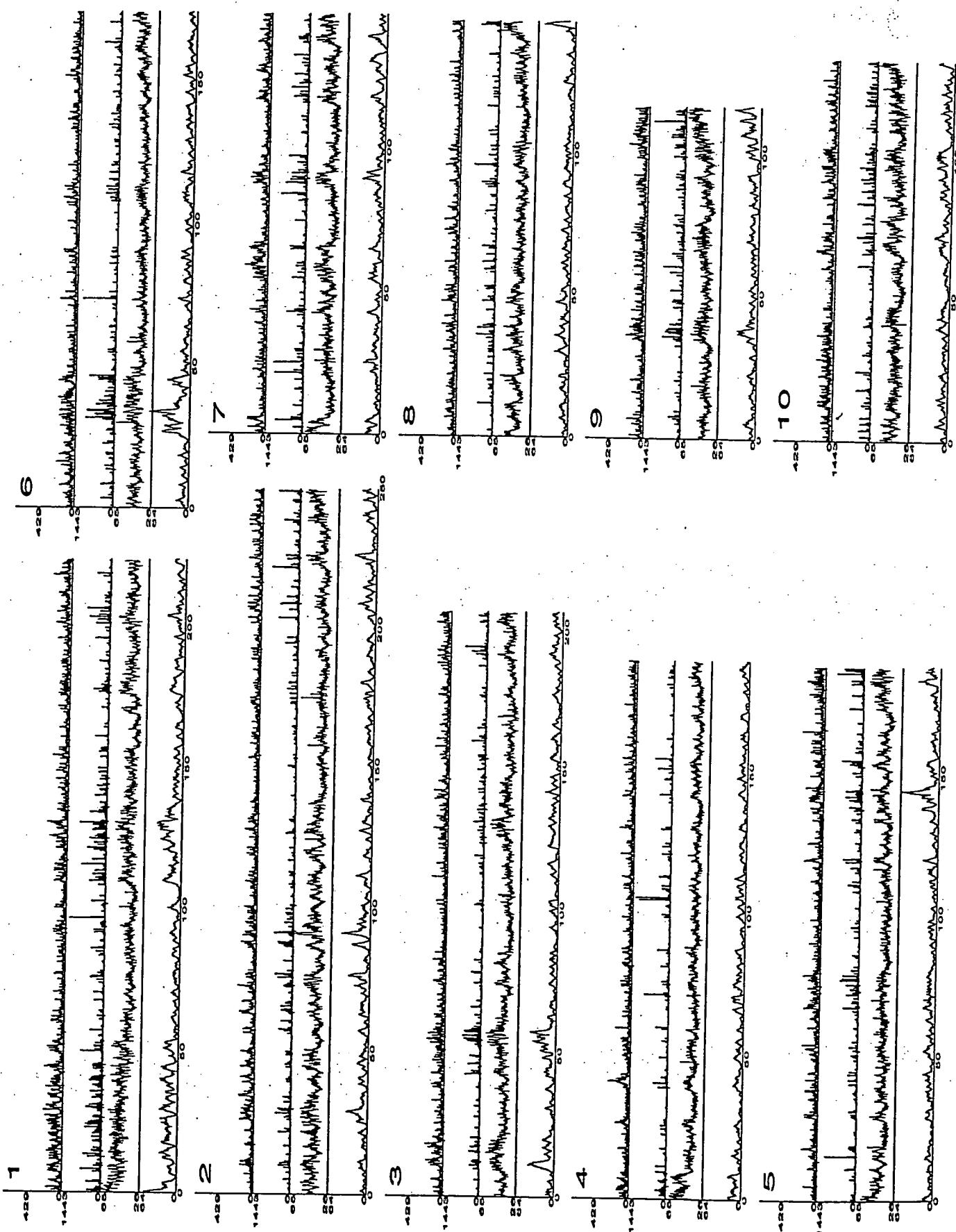


Fig. 11. Genome structural features.

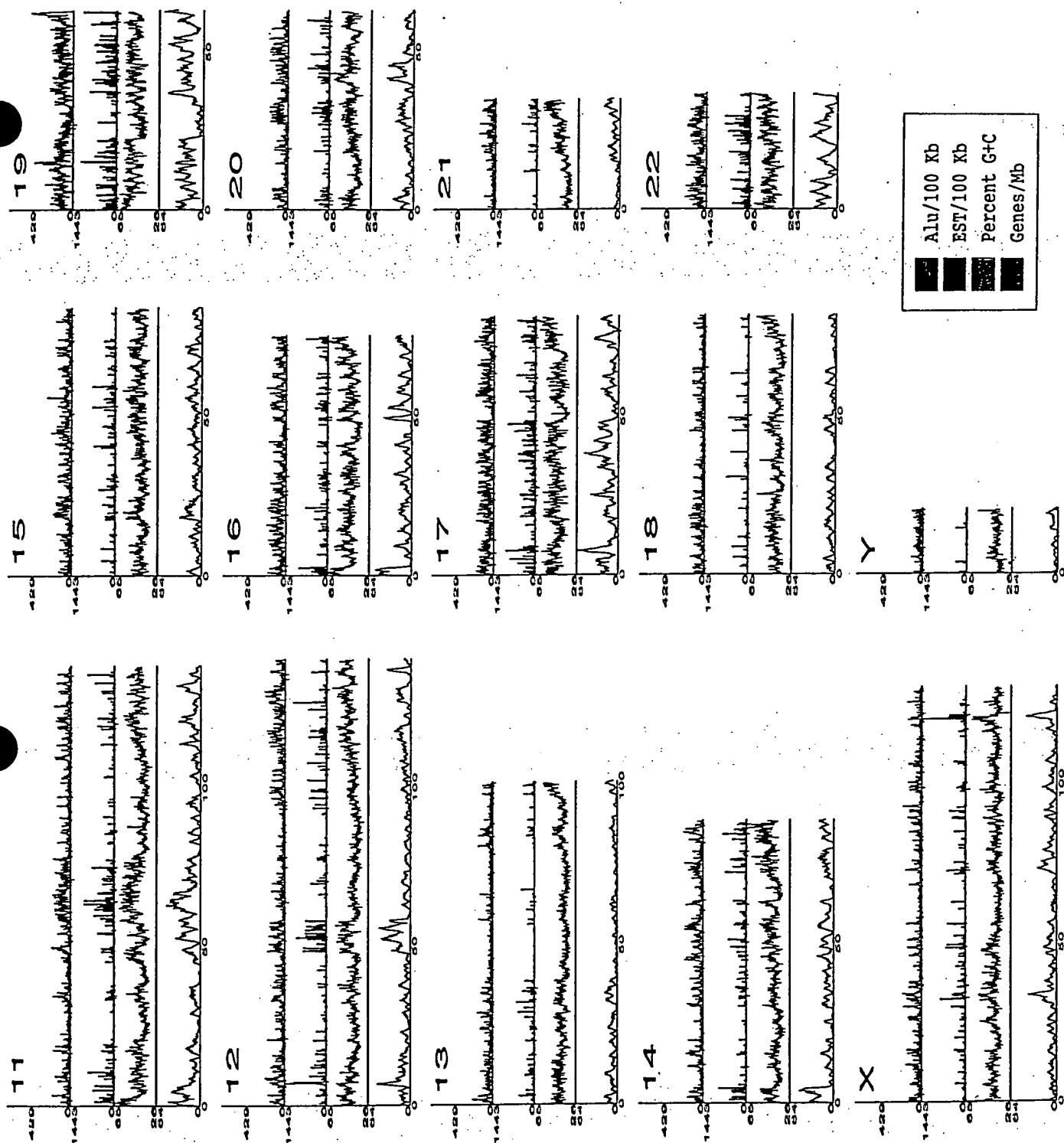


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows.

The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

### 5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of

a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (84, 85). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication

events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

## THE HUMAN GENOME

**Table 10.** Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

Sequence coverage (CCS assembly)				Base composition				Gene prediction*				Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaf-folds	Largest scaf-fold (Mbp)	No. of scaf-folds >500 kbp	% sequence covered by scaf-folds	% repeat	% GC	No. of CpG islands	Otto	De novo/any	De novo/2x	Total (Otto + de novo/any)	Sequence in deserts >500/1 kbp	Otto	De novo/any	De novo/2x	Otto + de novo/any	Otto + de novo/2x
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	29	6	8	3	16
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,554	55	19	5	7	11
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	50	12	5	7	7
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	55	18	4	6	8
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	46	15	5	7	6
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	38	9	6	7	11
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	26	12	5	7	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	22	9	6	7	6
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	27	9	5	7	13
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	24	9	8	8	12
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	34	20	7	8	11
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	3
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	7	14
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	10
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	15
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	4	7
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	7	10
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	3	7	11
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	9	6	12
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	0	14	8
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	5	6	11
U*	75	11,542	1	1,059	2,490	44	41	41	28,519	17,764	21,350	8,619	39,114	328	606	208	3	11
Total	2907	53,591																
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	2

\*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC1)) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (84, 86). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon-containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (87).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

## 5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not ex-

pressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

\*In these ranges, the percentages correspond to the annotated gene set (26,383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

## THE HUMAN GENOME

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3' end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed

pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (88). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG-non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

### 5.3 Gene duplication in the human genome

Building on a previously published procedure (27), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (89).

**Table 13.** Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of  $\geq 0.6$ . Method 2 uses a CG likelihood ratio of  $\geq 0.8$ .

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

**Table 14.** Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (83)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

### 5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family-based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were de-

termined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (89). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (89). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (90). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch -10, with gap open and extend penalties of -4 and -1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (91). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (92); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (93); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the

filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstruc-

tions at several evolutionary stages (94). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is  $2.3 \times 10^{-68}$  (93). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

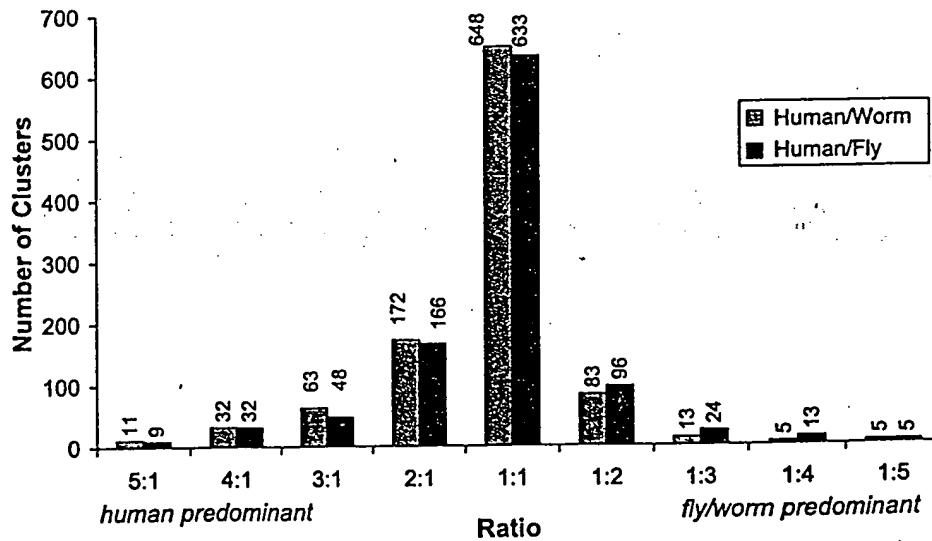


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (27). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

## THE HUMAN GENOME

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1](http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC1)). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a

pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (95). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (96). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually re-

veal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

### 6 A Genome-Wide Examination of Sequence Variations

**Summary.** Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (97), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (98). These data were consistent in showing an overall nucleotide diversity of  $\sim 8 \times 10^{-4}$ , marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

#### 6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (99). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (100) and in human SNPs.

(101, 102). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

## 6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP ([www.ncbi.nlm.nih.gov/SNP](http://www.ncbi.nlm.nih.gov/SNP)) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (103). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due

to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (46). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (101), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (46) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

## 6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used  $\pi$ , the standard statistic for nucleotide diversity (104). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced representation sequencing, we need to know the sequence quality and the depth of coverage at each

site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (105). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of  $\pi$  for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison,  $F = 29.73$ ,  $P < 0.0001$ ).

Average diversity for the autosomes estimated from the Celera-PFP comparison was  $8.94 \times 10^{-4}$ . Nucleotide diversity on the X chromosome was  $6.54 \times 10^{-4}$ . The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (106).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (101, 102, 106, 107). Genome-wide, our estimate of nucleotide diversity was  $8.98 \times 10^{-4}$  for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was  $8.00 \times 10^{-4}$  (108).

## 6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

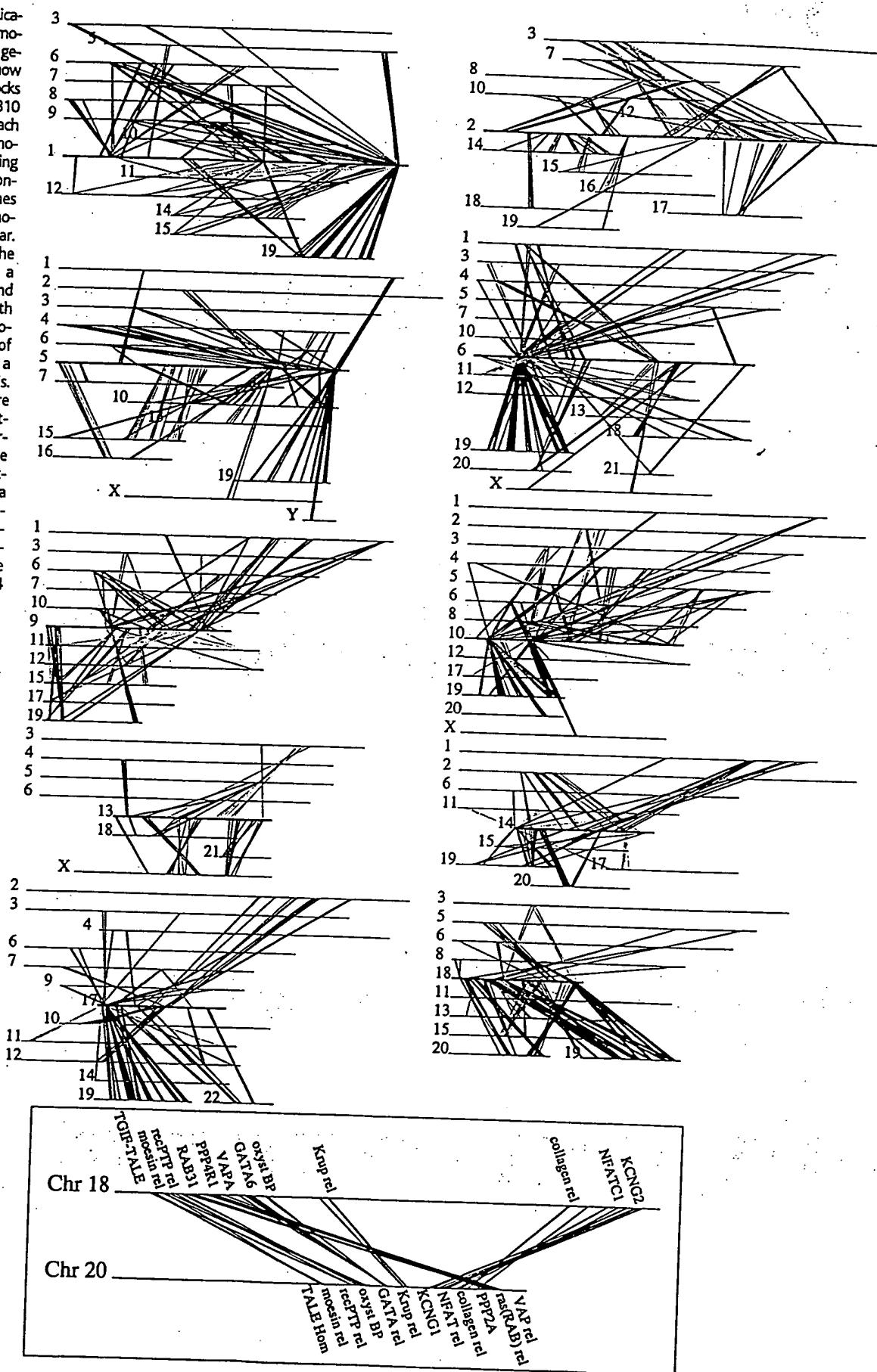
Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

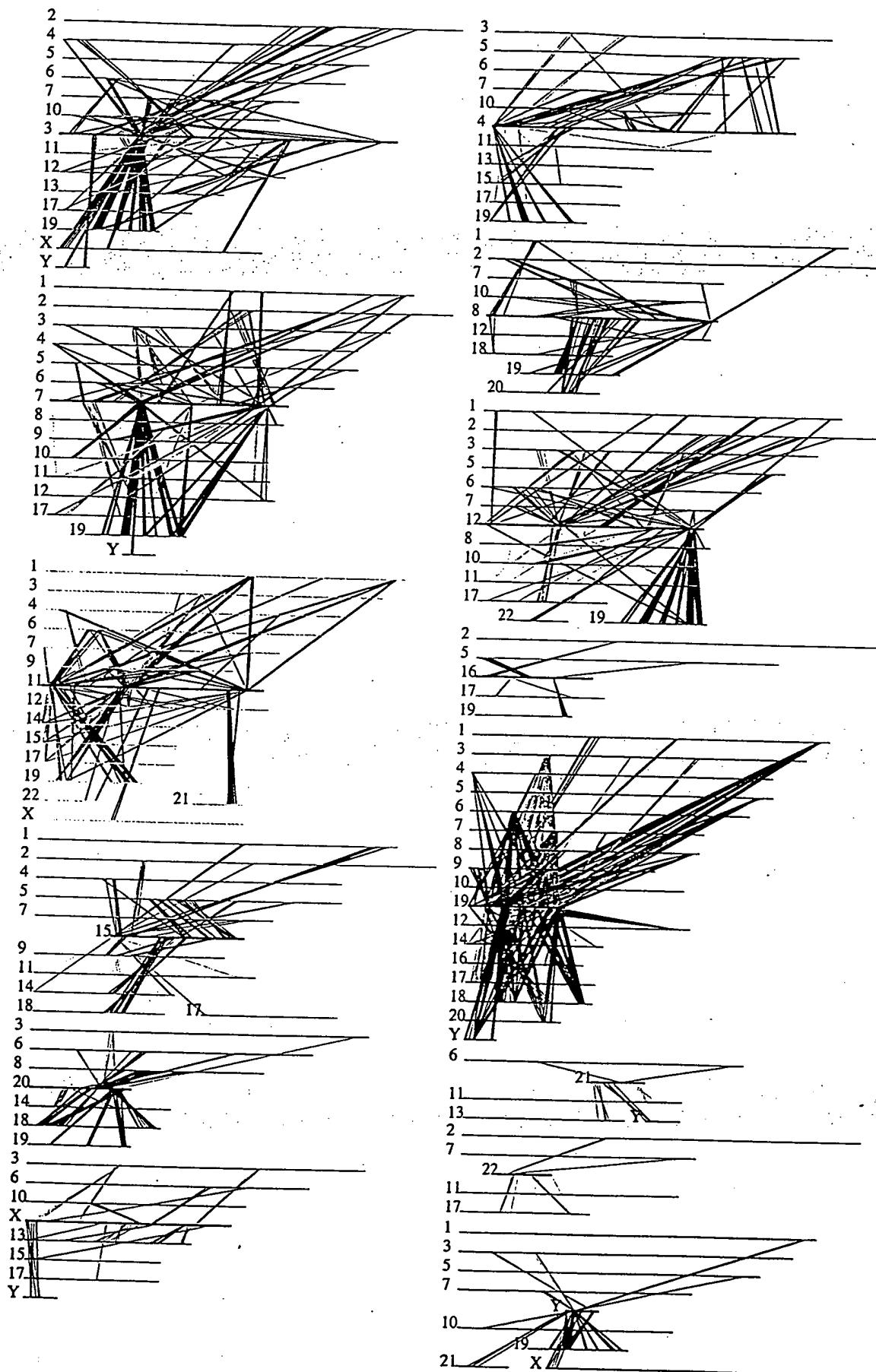
\*November 2000 release of the NCBI database dbSNP ([www.ncbi.nlm.nih.gov/SNP](http://www.ncbi.nlm.nih.gov/SNP)) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pu-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP ([www.ncbi.nlm.nih.gov/SNP](http://www.ncbi.nlm.nih.gov/SNP)) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

## THE HUMAN GENOME

**Fig. 13.** Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.



THE HUMAN GENOME



somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (109). Applying well-tested algorithms for simulating the neutral coalescent with recombination (110), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (111), we generated a distribution of numbers of SNPs by this model as well (112). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucle-

otides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive ( $r = 0.21$ ) and highly significant ( $P < 0.0001$ ), but G+C content accounted for only a small part of the variation.

### 6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as  $>5$  kbp from any predicted transcription unit), 5'-UTR, exonic (missense and silent), intronic, and 3'-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (112). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al.* (101) and by Halushka *et al.* (102). Similar results were observed in SNPs derived from Celera shotgun sequences (46).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about

0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (113), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (46). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

### 7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

**Summary.** This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain-based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (114, 115) and Celera's Panther Classification (CPC) (Fig. 15) (116). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (115, 117).

The results presented here are preliminary and are subject to several limitations.

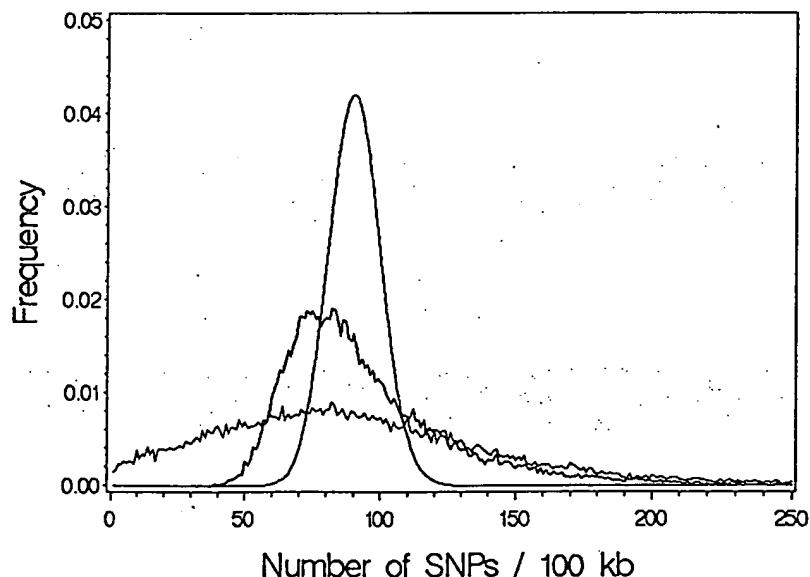


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals?

(iii) How does the human protein complement differ from that of other sequenced eukaryotes?

### 7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of

these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

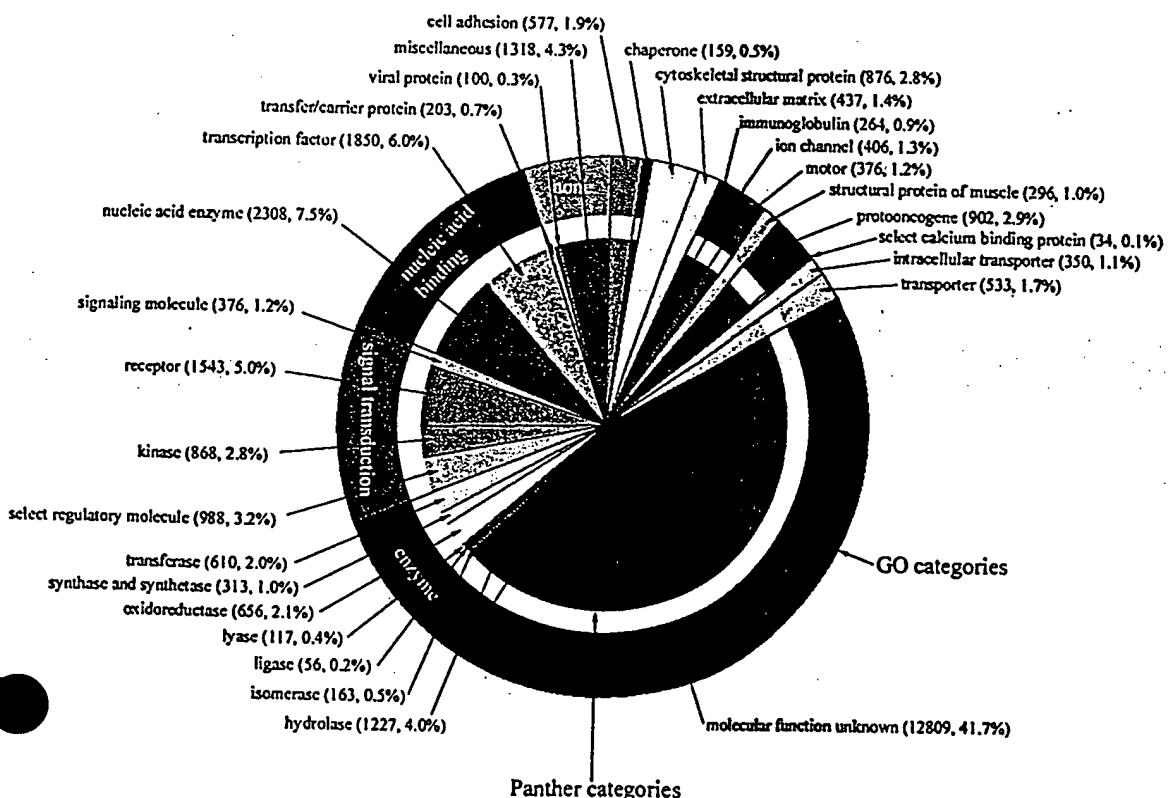


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (179), and the inner circle shows the assignment to Celera's Panther molecular function categories (116).

## THE HUMAN GENOME

### 7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (118) and two diverse invertebrates, *C. elegans* (a nematode worm) (119) and *D. melanogaster* (fly) (26), as well as the first plant genome, *A. thaliana*, recently completed (92), provide a diverse background for genome comparisons.

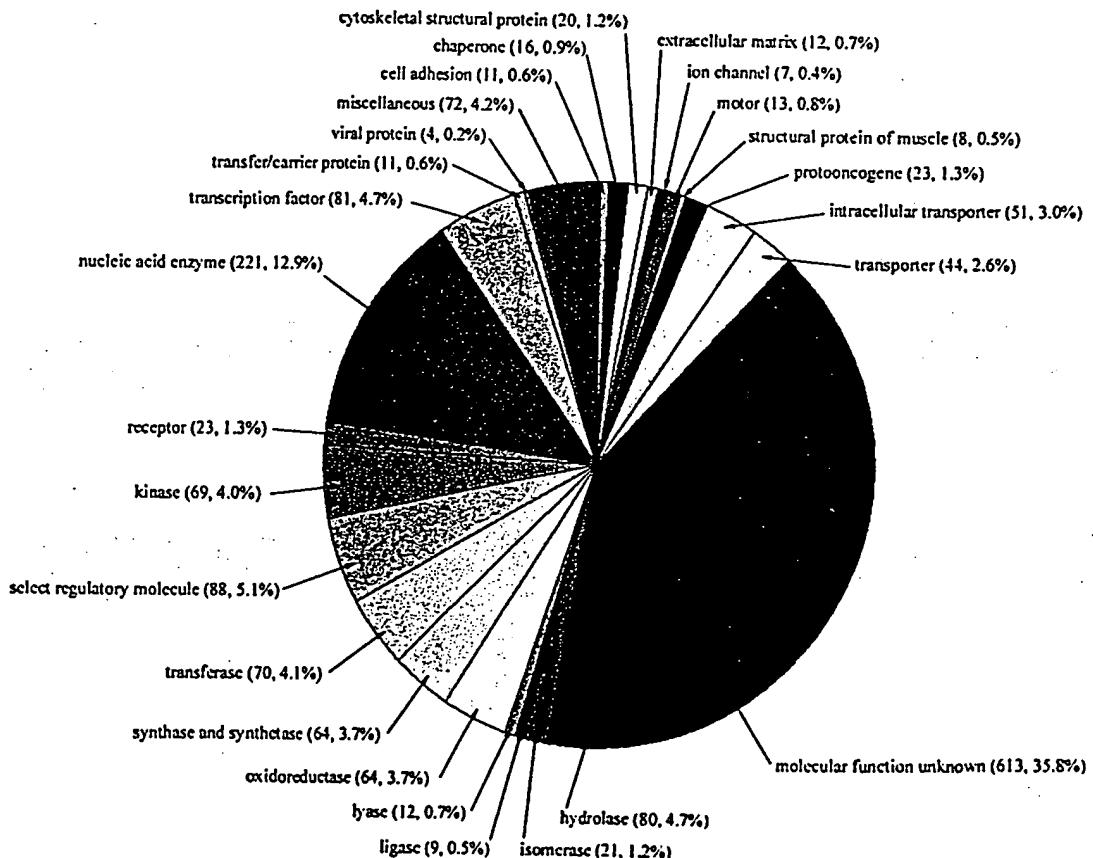
We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in

(120), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (120) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we con-

sider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans*.

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

**Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes.** Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (180) such that each orthologous pair (i) has a BLASTP *P*-value of  $\leq 10^{-10}$  (120), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).



zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

### 7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways

in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

**Acquired immunity.** One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

**Neural development, structure, and function.** In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (121); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number

of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain containing proteins (122) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (123). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (124). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (125). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (126).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a  $\text{Ca}^{2+}$  sensor (or receptor) during synaptic vesicle fusion and release (127). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (128). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (129).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycoprotein are found in the central nervous system. Mutations in any of these

# THE HUMAN GENOME

**Table 18.** Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in

more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (\*\*). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
<i>Developmental and homeostatic regulators</i>							
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	1	0	0	0	0
PF00029	Connexin	Connexin	3	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	14 (16)	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	1	0	0	0	0
PF00007	Cys_knot	Cystine-knot domain	2	1	0	0	0
PF00778	DIX	Dix domain	10 (11)	2	0	0	0
PF00322	Endothelin	Endothelin family	5	2	4	0	0
PF00812	Ephrin	Ephrin	3	0	0	0	0
PF01404	EPH_Ibd	Ephrin receptor ligand binding domain	7 (8)	2	4	0	0
PF00167	FGF	Fibroblast growth factor	12	2	1	0	0
PF01534	Frizzled	Frizzled/Smoothed family membrane region	23	1	1	0	0
PF00236	Hormone6	Glycoprotein hormones	9	7	3	0	0
PF01153	Glypican	Glypican	1	0	0	0	0
PF01271	Granin	Granin (chromogranin or secretogranin)	14	2	1	0	0
PF02058	Guanylin	Guanylin precursor	3	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	1	0	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	7	4	0	0	0
PF02024	Leptin	Leptin	10	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	1	0	0	0	0
PF00243	NGF	Nerve growth factor family	13 (23)	0	1	0	0
PF02158	Neuregulin	Neuregulin family	3	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	4	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	1	0	0	0	0
PF00865	Osteopontin	Osteopontin	3 (5)	2 (4)	2 (6)	0	0
PF00159	Hormone3	Pancreatic hormone peptides	1	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	3	0	0	0	0
PF00123	Hormone2	Peptide hormone	2	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5 (9)	0	0	0	0
PF01403	Sema	Sema domain	5	1	0	0	0
PF01033	Somatomedin_B	Somatomedin B domain	27 (29)	8 (10)	3 (4)	0	0
PF00103	Hormone	Somatotropin	5 (8)	3	0	0	0
PF02208	Sorb	Sorbin homologous domain	1	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	2	0	0	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	3	1	1	0	0
PF00019	TGF- $\beta$	Transforming growth factor $\beta$ -like domain	17 (31)	1	0	0	0
PF01099	Uteroglobin	Uteroglobin family	27 (28)	6	4	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	3	0	0	0	0
<i>Hemostasis</i>							
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal-like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (Continued)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptidé	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptidé	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	1	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
PI-PY-rho GTPase signaling							
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEF	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

## THE HUMAN GENOME

Table 18 (Continued)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
<i>Domains involved in apoptosis</i>							
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
<i>Cytoskeletal</i>							
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	0
PF01044	Vinculin	Vinculin family	4	2	1	0	5
<i>ECM adhesion</i>							
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
<i>Protein interaction domains</i>							
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	Dnaj_C	Dnaj C terminal region	12	9	5	3	19
PF00226	Dnaj	Dnaj domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

# THE HUMAN GENOME

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (130). Humans have at least 10 genes belonging to four different families involved in myelin produc-

tion (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (Continued)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans Isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
<i>Nuclear interaction domains</i>							
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRromatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5 (6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

## THE HUMAN GENOME

Table 18 (Continued)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	Zinc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor- $\beta$  (TGF- $\beta$ ), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (131). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (132), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (133). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and

homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain-containing proteins (134). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

**Hemostasis.** Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoietic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a

significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (135, 136). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor- $\alpha$ , and ADAM-10 has been implicated in the Notch signaling pathway (135). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

**Apoptosis.** Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (137). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (138). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain-containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

**Expansions of other protein families.** **Metabolic enzymes.** There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) are vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and they and their activators have been implicated

in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotrans-

posed GAPDH pseudogenes (139), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
<i>Neural structure, function, development</i>					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoreceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated $\text{Ca}^{2+}$ alpha	22	4	8	2	2
Voltage-gated $\text{Ca}^{2+}$ alpha-2	10	3	2	0	0
Voltage-gated $\text{Ca}^{2+}$ beta	5	2	2	0	0
Voltage-gated $\text{Ca}^{2+}$ gamma	1	0	0	0	0
Voltage-gated $\text{K}^{+}$ alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated $\text{Na}^{+}$	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
<i>Immune response</i>					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Interferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
SI cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor‡	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
Fl cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor§	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0

## THE HUMAN GENOME

alytic activity, as a uracil DNA glycosylase (140) and functions as a cell cycle regulator (141) and has even been implicated in apoptosis (142).

**Translation.** Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes

may account for many of these expansions [see the discussion above and (143)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (144).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retro-

transposition, and again there is evidence that many of these may be pseudogenes (145). However, a second form (eEF1A2) of this factor has been identified with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (146).

**Ribonucleoproteins.** Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

**Posttranslational modifications.** In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (147). The vitamin K-dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (148). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemo-kine receptors (149). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

**Concluding remarks.** There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, post-translational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential

Table 19 (Continued)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor-related	10	6	0	0	0
<i>Developmental and homeostatic regulators</i>					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/herculein	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoietin	2	0	1	0	0
Thyromosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
<i>Kinases and phosphatases</i>					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
<i>Signal transduction</i>					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

# THE HUMAN GENOME

Table 19 (Continued)

Panther family/subfamily*	H	F	W	Y	A
<i>Transcription factors/chromatin organization</i>					
C2H2 zinc finger-containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
<i>ECM adhesion</i>					
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	0
Integrin beta	9	2	2	0	1
LDL receptor family	26	19	20	0	0
Proteoglycans	22	9	7	0	5
<i>Apoptosis</i>					
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
<i>Hemostasis</i>					
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
<i>Other enzymes</i>					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
<i>Splicing and translation</i>					
EF-1alpha	56	13	10	6	13
Ribonucleaseprotein†	269	135	104	60	265
Ribosomal protein†	812	111	80	117	256

\*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

## THE HUMAN GENOME

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

### 8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3'- and 5'-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (153); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for

predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray-induced mutations at specific loci, Muller, in 1967 (154), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (155). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (156). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (26, 27). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (157); methylation of CpG islands in imprinting (158); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (159) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (160) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon

of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (161). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (162). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (163). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

*In situ* studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (68). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (69). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (71). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (164). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

### 8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplo-

types present in subjects of different ethno-geographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of intraspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele “swept” the linked variation out of the population (165). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (166). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

#### 8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of “this binds to that, which

then docks on this, and then the complex moves there. . . .” (167) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other “parts lists” reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (168). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (169), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about  $1.5 \text{ cm}^3$ , two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are “special cases” of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of “regulatory genes” that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF- $\beta$ , ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these “complexities” perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

#### 8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein’s brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (170). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (171). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (172), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no “good” genes or “bad” genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with “complexity,” particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human ge-

## THE HUMAN GENOME

ome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

### References and Notes

- R. L. Sinsheimer, *Genomics* 5, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
- R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
- F. Sanger et al., *Nature* 265, 687 (1977).
- P. H. Seeburg et al., *Trans. Assoc. Am. Physicians* 90, 109 (1977).
- E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* 154, 353 (1986).
- J. Gocayne et al., *Proc. Natl. Acad. Sci. U.S.A.* 84, 8296 (1987).
- A. Martin-Gallardo et al., *DNA Sequence* 3, 237 (1992); W. R. McCombie et al., *Nature Genet.* 1, 348 (1992); M. A. Jensen et al., *DNA Sequence* 1, 233 (1991).
- M. D. Adams et al., *Science* 252, 1651 (1991).
- M. D. Adams et al., *Nature* 355, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 373 (1993); M. H. Polymeropoulos et al., *Nature Genet.* 4, 381 (1993); M. Marra et al., *Nature Genet.* 21, 191 (1999).
- M. D. Adams et al., *Nature* 377, 3 (1995); O. White et al., *Nucleic Acids Res.* 21, 3829 (1993).
- F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* 162, 729 (1982).
- B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* 57, 577 (1991).
- R. D. Fleischmann et al., *Science* 269, 496 (1995).
- C. M. Fraser et al., *Science* 270, 397 (1995).
- C. J. Bult et al., *Science* 273, 1058 (1996); J. F. Tomb et al., *Nature* 388, 539 (1997); H. P. Klenk et al., *Nature* 390, 364 (1997).
- J. C. Venter, H. O. Smith, L. Hood, *Nature* 381, 364 (1996).
- H. Schmitt et al., *Genomics* 33, 9 (1996).
- S. Zhao et al., *Genomics* 63, 321 (2000).
- X. Lin et al., *Nature* 402, 761 (1999).
- J. L. Weber, E. W. Myers, *Genome Res.* 7, 401 (1997).
- P. Green, *Genome Res.* 7, 410 (1997).
- E. Pennisi, *Science* 280, 1185 (1998).
- J. C. Venter et al., *Science* 280, 1540 (1998).
- M. D. Adams et al., *Nature* 368, 474 (1994).
- E. Marshall, E. Pennisi, *Science* 280, 994 (1998).
- M. D. Adams et al., *Science* 287, 2185 (2000).
- G. M. Rubin et al., *Science* 287, 2204 (2000).
- E. W. Myers et al., *Science* 287, 2196 (2000).
- F. S. Collins et al., *Science* 282, 682 (1998).
- International Human Genome Sequencing Consortium (2001), *Nature* 409, 860 (2001).
- Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
- Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
- DNA was isolated from blood (173) or sperm. For sperm, a washed pellet (100  $\mu$ l) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl-20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to *Bst* XI adaptors (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adaptors, and the fragments, now with 3'-CACA overhangs, were inserted into *Bst* XI-linearized plasmid vector with 3'-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with *Bgl* II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp *Bam* HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of *Bgl* II. As *Bgl* II-*Bgl* II ligations occurred, they were continually cleaved, whereas *Bam* HI-*Bgl* II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50  $\mu$ g/ml), carbenicillin (50  $\mu$ g/ml), and kanamycin (15  $\mu$ g/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
- Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (173) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

# THE HUMAN GENOME

sample plate barcode, thus enhancing sample sheet-to-plate associations.

35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977); J. M. Prober *et al.*, *Science* **238**, 336 (1987).

36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).

37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (174), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5' vector splice junction in the initial part of a read was removed.

38. National Center for Biotechnology Information (NCBI); available at [www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/).

39. NCBI; available at [www.ncbi.nlm.nih.gov/HTGS/](http://www.ncbi.nlm.nih.gov/HTGS/).

40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.

41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (175), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.

42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996).

43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.

44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).

45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).

46. J. Zhang *et al.*, data not shown.

47. Shaded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.

48. M. Hattori *et al.*, *Nature* **405**, 311 (2000).

49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).

50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).

51. The International RH Mapping Consortium, available at [www.ncbi.nlm.nih.gov/genemap99/](http://www.ncbi.nlm.nih.gov/genemap99/).

52. See <http://ftp.genome.washington.edu/RM/RepeatMasker.html>.

53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998).

54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).

55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).

55b. See <http://genome.ucsc.edu/>.

56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).

57. D. Dickson, *Nature* **401**, 311 (1999).

58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).

59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).

60. M. Yandell, in preparation.

61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).

62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (52) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3X), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (54) optimized for the Compaq Alpha computer-server and an effective database size of  $3 \times 10^9$  for BLASTN searches and  $1 \times 10^9$  for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of  $<1 \times 10^{-4}$ , human nucleotide BLAST results having an expectation score of  $<1 \times 10^{-8}$  with >94% identity, and rodent nucleotide BLAST results having an expectation score of  $<1 \times 10^{-8}$  with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (63).

63. E. C. Überbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).

64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).

65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.

66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).

67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).

68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).

69. G. Bernardi, *Gene* **241**, 3 (2000).

70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).

71. S. Ohno, *Trends Genet.* **1**, 160 (1985).

72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998).

73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).

74. A. Bird, *Trends Genet.* **3**, 342 (1987).

75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).

76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).

77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).

78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (<http://genomebiology.com/2000/1/5/reviews/1028>).

79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000).

80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).

81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).

82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).

83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).

84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).

85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).

86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).

87. P. Nouvel, *Genetica* **93**, 191 (1994).

88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).

89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair  $ij$  in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by  $i$  and  $j$ . This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B-containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

## THE HUMAN GENOME

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (27) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: if one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).

91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).

92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).

93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A' and B' (paralogs of A and B) also exist in the genome, the probability that B' occurs immediately after A' is  $1/N$ , where  $N$  is the number of proteins in the set (for this analysis,  $N = 26,588$ ). Allowing for B' to occur as any of the next  $J-1$  proteins [leaving a gap between A' and B' increases the probability to  $(J-1)/N$ ; allowing B'A' or A'B' gives a probability of  $2(J-1)/N$ ]. Considering three genes ABC, the probability of observing A'B'C' elsewhere in the genome, given that the paralogs exist, is  $1/N^2$ . Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that  $K$  proteins can be spread across  $J$  positions by counting all possible arrangements of  $K-2$  proteins in the  $J-2$  positions between the first and last protein. Allowing for a spread to vary from  $K$  positions (no gaps) to  $J$  gives

$$L = \sum_{x=k-2}^{J-2} \binom{X}{K-2}$$

arrangements. Thus, the probability of chance occurrence is  $L/N^{K-1}$ . Allowing for both sets of genes (e.g., ABC and A'B'C') to be spread across  $J$  positions increases this to  $L^2/N^{K-1}$ . The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for  $M$  such rearrangements gives us a probability  $P = L^2 M / N^{K-1}$ . For example, the

probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is  $36/N^2$ ; the expected number of such matched sets in the predicted protein set is approximately  $(N)36/N^2 = 36/N$ , a value  $\ll 1$ . Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with  $P \ll 1$ .

94. B. J. Trask *et al.*, *Hum. Mol. Genet.* **7**, 13 (1998); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. Mclysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marti *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage  $x$  from a given individual, both homologs are present in the assembly with probability  $1 - (1/2)^{x-1}$ . Even if both homologs are present, the probability that a SNP is detected is  $< 1$  because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.
106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* **10**, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavaré, *Theor. Popul. Biol.* **26**, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1-44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value  $< 10^{-3}$ ) and "globally" alignable (the length of the match region must be  $> 70\%$  and  $< 130\%$  of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NULL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (176) were then used as input into the BETE program (177). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (178) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NULL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.
117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999).
118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessel, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonelli, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999).

## THE HUMAN GENOME

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998).

137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).

138. A. G. Uren et al., *Mol. Cell* **6**, 961 (2000).

139. P. Garcia-Meunier, M. Etienne-Julian, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993).

140. K. Meyer-Siegler et al., *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).

141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993).

142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000).

143. N. Kenmochi et al., *Genome Res.* **8**, 509 (1998).

144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999).

145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990).

146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999).

147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000).

148. P. Munroe et al., *Nature Genet.* **21**, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991); B. Furie et al., *Blood* **93**, 1798 (1999).

149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000).

150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).

151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999).

152. C. M. Fraser et al., *Science* **281**, 375 (1998); H. Tettelin et al., *Science* **287**, 1809 (2000).

153. D. Brett et al., *FEBS Lett.* **474**, 83 (2000).

154. H. J. Muller, H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967).

155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.

156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).

157. K. Kobayashi et al., *Nature* **394**, 388 (1998).

158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000).

159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000).

160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).

161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).

162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000).

163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000).

164. E. Capanha, M. G. M. Romanini, *Caryologia* **24**, 471 (1971).

165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987).

166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).

167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999).

168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998).

169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993).

170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* **2**, 44 (1996).

171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).

172. E. Colucci-Guyon et al., *Cell* **79**, 679 (1994).

173. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).

174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).

175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).

176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).

177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).

178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).

179. GO, available at [www.geneontology.org/](http://www.geneontology.org/).

180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site ([www.celera.com](http://www.celera.com)). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001

# Science

## Functional Genomics Web Site

- Links to breaking news in genomics and biotech from *Science*, *ScienceNOW*, and other sources
- Pointers to classic papers, news, and new research, organized by subject, relevant to the post-genomics world
- *Science's* genome special issues
- Collections of Web resources in genomics and post-genomics, including special pages on model organisms, educational resources, and gene maps.
- A central node of news, information, and links to the biotech business.

[www.sciencegenomics.org](http://www.sciencegenomics.org)

# THE HUMAN GENOME



Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two

ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at [www.sciencemag.org/feature/data/announcement/gsp.shl](http://www.sciencemag.org/feature/data/announcement/gsp.shl).) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere: Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

A historic  
moment for  
the scientific  
endeavor.

Barbara R. Jasny and Donald Kennedy

Blast SEQ ID NO:1 versus human genomic

MEGABLAST 1.2.3-Paracel [2001-11-20]

Reference:

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000),  
"A greedy algorithm for aligning DNA sequences",  
J Comput Biol 2000; 7(1-2):203-14.

Database: Homo\_sapiens.latestgp.masked.fa  
44,521 sequences; 200,768,834,160 total letters

Query= LEX161seqid1  
(2115 letters)

Value	Score (bits)	E
AC016643.6.1.179969	734	0.0
AC008958.7.1.109526	520	e-
144		

Sequences producing significant alignments:

>AC016643.6.1.179969  
Length = 179969

Score = 734 bits (370), Expect = 0.0  
Identities = 370/370 (100%)  
Strand = Plus / Plus

Query: 957 agatgatttaagccagaagttaaaacccttgggtgaagcagaacgagagttatggaa  
1016 |||||||  
Sbjct: 60550 agatgatttaagccagaagttaaaacccttgggtgaagcagaacgagagttatggaa  
60609

Query: 1017 ttgaagaaaaaggaatgcaaagacaggggtttgaatatgtggaaaatcaatgcctg  
1076 |||||||  
Sbjct: 60610 ttgaagaaaaaggaatgcaaagacaggggtttgaatatgtggaaaatcaatgcctg  
60669

Query: 1077 ggatctatattactacatgactcagacagaggaactcaagtattccatagaccaaggtt  
1136 |||||||  
Sbjct: 60670 ggatctatattactacatgactcagacagaggaactcaagtattccatagaccaaggtt  
60729

Query: 1137 cctcaaggaatacttcccaattgaggtggtcactgaaggcttgcgaacacacctaccagga  
1196 |||||||  
Sbjct: 60730 cctcaaggaatacttcccaattgaggtggtcactgaaggcttgcgaacacacctaccagga  
60789

Query: 1197 gttgttggacttcattgaacaaatgacagatgctcatgttggaaacaagagtgtac  
1256 |||||||  
Sbjct: 60790 gttgttggacttcattgaacaaatgacagatgctcatgttggaaacaagagtgtac  
60849

Query: 1257 actttatactgtgaaggataaagctacaggagaagtattggacagttctattggacct  
1316 |||||||  
Sbjct: 60850 actttatactgtgaaggataaagctacaggagaagtattggacagttctattggacct  
60909

Query: 1317 ctatccaagg 1326  
|||||||  
Sbjct: 60910 ctatccaagg 60919

Score = 520 bits (262), Expect = e-144  
Identities = 262/262 (100%)  
Strand = Plus / Plus

Query: 40 agagttgggtgtccaggatttactcagaatgacgttaggaagagaagtgtatgtctccct  
99 |||||||  
Sbjct: 30994 agagttgggtgtccaggatttactcagaatgacgttaggaagagaagtgtatgtctccct  
31053

Query: 100 cttcaggcaatgtcttctatactgtggctggcagaaatgtttaagatggatcttca  
159 |||||||  
Sbjct: 31054 cttcaggcaatgtcttctatactgtggctggcagaaatgtttaagatggatcttca  
31113

Query: 160 ccagagcaaattaaaacaagaactgaggagctattgtcagaccaaacaggtgtacgt  
219 |||||||  
Sbjct: 31114 ccagagcaaattaaaacaagaactgaggagctattgtcagaccaaacaggtgtacgt  
31173

Query: 220 gctgttggaatgctcggtattgaggaacttacgagaactgtctgcaggcactggca  
279 |||||||  
Sbjct: 31174 gctgttggaatgctcggtattgaggaacttacgagaactgtctgcaggcactggca  
31233

Query: 280 gatgtagaagtaaaagtatata 301  
|||||||  
Sbjct: 31234 gatgtagaagtaaaagtatata 31255

Score = 405 bits (204), Expect = e-110

Identities = 204/204 (100%)  
Strand = Plus / Plus

Query: 1324 aggaaaggaaaatacatgcggcctgttcggtccagcctggctgccttcgcct  
1383 |||||||

Sbjct: 64886 aggaaaggaaaatacatgcggcctgttcggtccagcctggctgccttcgcct  
64945 |||||||

Query: 1384 gatggaagccggatgatggcagtggctgcctcgtaacttcacagccagtggca  
1443 |||||||

Sbjct: 64946 gatggaagccggatgatggcagtggctgcctcgtaacttcacagccagtggca  
65005 |||||||

Query: 1444 ggtcgccctctcctgagacacgacgaggtgaggacttacttcatgagttggcac  
1503 |||||||

Sbjct: 65006 ggtcgccctctcctgagacacgacgaggtgaggacttacttcatgagttggcac  
65065 |||||||

Query: 1504 gtgatgcatcagattgtgcacag 1527  
|||

Sbjct: 65066 gtgatgcatcagattgtgcacag 65089

Score = 379 bits (191), Expect = e-102

Identities = 191/191 (100%)  
Strand = Plus / Plus

Query: 1526 agactgatttgcacgatttagcgaaacaaatgtggaaactgacttttagaggtgccat  
1585 |||||||

Sbjct: 81947 agactgatttgcacgatttagcgaaacaaatgtggaaactgacttttagaggtgccat  
82006 |||||||

Query: 1586 cgcaaatgctgaaaattgggtgtggacgtcgattccctccgaagattgtcaaaacatt  
1645 |||||||

Sbjct: 82007 cgcaaatgctgaaaattgggtgtggacgtcgattccctccgaagattgtcaaaacatt  
82066 |||||||

Query: 1646 ataaaagatggaagccctattgcagacgatctgctgaaaaacttggctttaggctgg  
1705 |||||||

Sbjct: 82067 ataaaagatggaagccctattgcagacgatctgctgaaaaacttggctttaggctgg  
82126 |||||||

Query: 1706 tcaacacaggt 1716  
|||

Sbjct: 82127 tcaacacaggt 82137

Score = 323 bits (163), Expect = 2e-85  
Identities = 163/163 (100%)  
Strand = Plus / Plus

Query: 661 ggtgcttctgtatgtttcattgacagtttagaaaagacagatgtgacaagtataaa  
720 |||||||  
Sbjct: 53694 ggtgcttctgtatgtttcattgacagtttagaaaagacagatgtgacaagtataaa  
53753

Query: 721 attacctaaaatccacactattccctgtcatgaagaaatgttgtatccctgaaacc  
780 |||||||  
Sbjct: 53754 attacctaaaatccacactattccctgtcatgaagaaatgttgtatccctgaaacc  
53813

Query: 781 agaagaaggatggaaatggcttttaatacacaagggtgcaaagagg 823  
|||  
Sbjct: 53814 agaagaaggatggcttttaatacacaagggtgcaaagagg 53856

Score = 301 bits (152), Expect = 8e-79  
Identities = 152/152 (100%)  
Strand = Plus / Plus

Query: 300 agtggaaaggaccatgtactttcccccagcatgtatcctctgacaaagaagtacgagc  
359 |||||||  
Sbjct: 35386 agtggaaaggaccatgtactttcccccagcatgtatcctctgacaaagaagtacgagc  
35445

Query: 360 agcaagtacagaaggcagacaaaagactttctgtttgatattgagatgagcatgagagg  
419 |||||||  
Sbjct: 35446 agcaagtacagaaggcagacaaaagactttctgtttgatattgagatgagcatgagagg  
35505

Query: 420 agatataattgagagaattgttcatttacagg 451  
|||  
Sbjct: 35506 agatataattgagagaattgttcatttacagg 35537

Score = 280 bits (141), Expect = 3e-72  
Identities = 141/141 (100%)  
Strand = Plus / Plus

Query: 1842 aggcacaaaatgccagctacccgttggacattggcaggggatacgatggccaatatta  
1901 |||||||

Sbjct: 84687 aggcacaaatgccagctacccatggacattggcaggggatacgatggccaatatta  
84746

Query: 1902 tggatatcttggagtgaagtatttccatggatatgtttacagctgtttaaaaaaga  
1961 |||||||

Sbjct: 84747 tggatatcttggagtgaagtatttccatggatatgtttacagctgtttaaaaaaga  
84806 |||||||

Query: 1962 agggataatgaatccagaggt 1982

Sbjct: 84807 agggataatgaatccagaggt 84827

Score = 270 bits (136), Expect = 3e-69

Identities = 137/138 (99%)

Strand = Plus / Plus

Query: 821 agaaaaacaccataatttgcagcagctactccactgcgaaccaagggtggccaaactac  
880 |||||||

Sbjct: 58207 agaaaaacaccataatttgcagcagctactccactgcgaaccaagggtggccaaactac  
58266 |||||||

Query: 881 tcggttatagcacacatgctgacttcgtccttcaaatgaacactgcacaaagcacaagcc  
940 |||||||

Sbjct: 58267 tcggttatagcacacatgctgacttcgtccttcaaatgaacactgcacaaagcacaagcc  
58326 |||||||

Query: 941 gcgtaacagcyttctag 958

Sbjct: 58327 gcgtaacagccttctag 58344

Score = 268 bits (135), Expect = 1e-68

Identities = 136/137 (99%)

Strand = Plus / Plus

Query: 1979 agtttggaaatgaaatacagaaacctaattcctgaaacctggggatctctggacggcatgg  
2038 |||||||

Sbjct: 95214 agtttggaaatgaaatacagaaacctaattcctgaaacctggggatctctggacggcatgg  
95273 |||||||

Query: 2039 acatgctccacaatttcttggaaacgtgagccaaaccaaaaaagcgttcctaattgagtagag  
2098 |||||||

Sbjct: 95274 acatgctccacaatttcttggaaacgtgagccaaaccaaaaaagcgttcctaattgagtagag  
95333 |||||||

Query: 2099 gcctgcatgctycgtga 2115  
||||||| |||||  
Sbjct: 95334 gcctgcatgctccgtga 95350

Score = 262 bits (132), Expect = 7e-67  
Identities = 132/132 (100%)  
Strand = Plus / Plus

Query: 1713 aggtcttctgaccctgcgccagattgtttgagcaaagttgatcagtctcttcataccaa  
1772 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 82469 aggtcttctgaccctgcgccagattgtttgagcaaagttgatcagtctcttcataccaa  
82528

Query: 1773 cacatcgctggatgctgcaagtgaatatgccaaatactgctcagaaatattaggagttgc  
1832 ||||||||||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 82529 cacatcgctggatgctgcaagtgaatatgccaaatactgctcagaaatattaggagttgc  
82588

Query: 1833 agctactccagg 1844  
|||||||  
Sbjct: 82589 agctactccagg 82600

Score = 222 bits (112), Expect = 6e-55  
Identities = 116/117 (99%), Gaps = 1/117 (0%)  
Strand = Plus / Plus

Query: 443 atttacagggaaacctgtatctggggaaagataaaacctgaggccagacgatacttggaaa  
502 ||||| ||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 49854 attt-cagggaaacctgtatctggggaaagataaaacctgaggccagacgatacttggaaa  
49912

Query: 503 agtcaattaaaatggggaaaagaaaatgggctccatctcctgaacaagtacagaatg 559  
|||||||  
Sbjct: 49913 agtcaattaaaatggggaaaagaaaatgggctccatctcctgaacaagtacagaatg  
49969

Score = 208 bits (105), Expect = 1e-50  
Identities = 105/105 (100%)  
Strand = Plus / Plus

Query: 559 gaaatcaaataatgaagaaaagaatgagtgagctatgtattgatttaacaaaaacctc  
618 ||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 52613 gaaatcaaataatgaagaaaagaatgagtgagctatgtattgatttaacaaaaacctc  
52672

Query: 619    aatgaggatgataccttccttgtatccaaaggctgaacttggg 663  
          |||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 52673 aatgaggatgataccttccttgtatccaaaggctgaacttggg 52717

>AC008958.7.1.109526  
Length = 109526  
Score = 520 bits (262), Expect = e-144  
Identities = 262/262 (100%)  
Strand = Plus / Minus

Query: 40    agagttggtggttccaggatttactcagaatgacgttaggaagagaagtgtatgtctcct  
99           |||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 24199 agagttggtggttccaggatttactcagaatgacgttaggaagagaagtgtatgtctcct  
24140

Query: 100    cttcaggcaatgtcttcctatactgtggctggcagaaatgtttaaagatggatcttc  
159           |||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 24139 cttcaggcaatgtcttcctatactgtggctggcagaaatgtttaaagatggatcttc  
24080

Query: 160    ccagagcaaattaaaacaagaactgaggagctcattgtgcagaccaaacagggtgtacat  
219           |||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 24079 ccagagcaaattaaaacaagaactgaggagctcattgtgcagaccaaacagggtgtacat  
24020

Query: 220    gctgttggaatgctcggtattgaggaagtaacttacgagaactgtctgcaggcactggca  
279           |||||||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 24019 gctgttggaatgctcggtattgaggaagtaacttacgagaactgtctgcaggcactggca  
23960

Query: 280    gatgtagaagtaaaagtatata 301  
          |||||||||||||||||||  
Sbjct: 23959 gatgtagaagtaaaagtatata 23938

Score = 323 bits (163), Expect = 2e-85  
Identities = 163/163 (100%)  
Strand = Plus / Minus

Query: 661    ggtgcttcctgtatgttgcacgttttagaaaagacagatgtacaagtataaa  
720           |||||||||||||||||||||||||||||||||||||||||||  
Sbjct: 1504 ggtgcttcctgtatgttgcacgttttagaaaagacagatgtacaagtataaa  
1445

Query: 721 attacctaaaatatccacactattccctgtcatgaagaaatgtgtatccctgaaacc  
780 |||||||  
Sbjct: 1444 attacctaaaatatccacactattccctgtcatgaagaaatgtgtatccctgaaacc  
1385

Query: 781 agaagaaggatggaaatggctttaatacaaggtgcaaagagg 823  
Sbjct: 1384 agaagaaggatggaaatggctttaatacaaggtgcaaagagg 1342

Score = 301 bits (152), Expect = 8e-79  
Identities = 152/152 (100%)  
Strand = Plus / Minus

Query: 300 agtggaaaggaccatgctagactttccccagcatgtatcctctgacaaagaagtacgagc  
359 |||||||  
Sbjct: 19806 agtggaaaggaccatgctagactttccccagcatgtatcctctgacaaagaagtacgagc  
19747

Query: 360 agcaagtacagaaggcagacaaaagactttctcgaaaaatggatggatggatggatggagg  
419 |||||||  
Sbjct: 19746 agcaagtacagaaggcagacaaaagactttctcgaaaaatggatggatggatggatggagg  
19687

Query: 420 agatataattgagagaattgtcattacagg 451  
Sbjct: 19686 agatataattgagagaattgtcattacagg 19655

Score = 222 bits (112), Expect = 6e-55  
Identities = 116/117 (99%), Gaps = 1/117 (0%)  
Strand = Plus / Minus

Query: 443 atttacagggaaacctgtatctgggaagataaaacctgaggccagacgatacttggaaa  
502 |||||||  
Sbjct: 5344 attt-cagggaaacctgtatctgggaagataaaacctgaggccagacgatacttggaaa  
5286

Score = 208 bits (105), Expect = 1e-50  
Identities = 105/105 (100%)  
Strand = Plus / Minus

Query: 559 gaaatcaaataatgaagaaaagaatgagtgagctatgtattgatttaacaaaacctc  
618 |||||||  
Sbjct: 2585 gaaatcaaataatgaagaaaagaatgagtgagctatgtattgatttaacaaaacctc  
2526

Query: 619 aatgaggatgataccttccttgcattttcaaggctgaacttgg 663  
|||||||  
Sbjct: 2525 aatgaggatgataccttccttgcattttcaaggctgaacttgg 2481

Score = 81.8 bits (41), Expect = 2e-12  
Identities = 41/41 (100%)  
Strand = Plus / Minus

Query: 1 atgatcgcccggtgcctttggctgtgcgaaggcctccgcag 41  
|||||||  
Sbjct: 60390 atgatcgcccggtgcctttggctgtgcgaaggcctccgcag 60350